Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Sistema experto probabilístico basado en redes bayesianas para la predicción de riesgo de cáncer cervical

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Luis Alonso PAULINO FLORES

ASESOR

Ing. Ana María HUAYNA DUEÑAS

Lima, Perú

2019

Referencia bibliográfica

Paulino, L. (2019). *Sistema experto probabilístico basado en redes bayesianas para la predicción de riesgo de cáncer cervical.* [Tesis de Ingeniería de Sistemas, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

HOJA DE METADATOS COMPLEMENTARIOS

Código ORCID del autor	—
DNI o pasaporte del autor	72468049
Código ORCID del asesor	https://orcid.org/0000-0001-7726-8206
DNI o pasaporte del asesor	06017183
Grupo de investigación	Inteligencia Artificial
Agencia financiadora	NO
Ubicación geográfica donde se desarrolló la investigación	Perú, Lima, Lima, San Miguel Latitud: -12.076161 | Longitud: -77.099655
Año ó rango de años en que se realizó la investigación	2017 - 2019
Disciplinas OCDE	Ingeniería de sistemas y comunicaciones http://purl.org/pe-repo/ocde/ford#2.02.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta de Sustentación de Tesis

Siendo las..6:20.....horas del día..19..de noviembre del año 2019 se reunieron los docentes designados como miembros de Jurado de Tesis, presidido por el Dr. Hugo Froilán Vega Huerta (Presidente), la Mg. Virginia Vera Pomalaza (Miembro) y la Ing. Ana María Huayna Dueñas (Miembro Asesor) para la sustentación de la Tesis Intitulada: **"SISTEMA EXPERTO PROBABILÍSTICO BASADO EN REDES BAYESIANAS PARA LA PREDICCIÓN DE RIESGO DE CÁNCER CERVICAL",** del Bachiller: **Luis Alonso Paulino Flores;** para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

El Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de.....19..... (En letras)..DIECINUEVE.

A continuación el Presidente del Jurado Dr. Hugo Froilán Vega Huerta, declara al Bachiller **Ingeniero de Sistemas.**

Siendo las..7:10.....horas, se levantó la sesión.

.....
Presidente
Dr. Hugo Froilán Vega Huerta

.....
Miembro
Mg. Virginia Vera Pomalaza

.....
Miembro Asesor
Ing. Ana María Huayna Dueñas

DEDICATORIA

A mi mamá que desde el cielo siempre me cuida y me acompaña en cada paso que doy. A mi papá por su apoyo incondicional, su amor, su confianza y sus grandes consejos.

AGREDECIMIENTOS

A mi asesora, profesora Ana María, por su acertada orientación y motivación constante.

Al Hospital Militar de la Fuerza Aérea del Perú; al área de Educación y de Registros Hospitalarios por aprobar el proyecto; al área de Ginecología y Obstetricia, especialmente a la Dra. Carmen Quispe y la enfermera Kathy Oros por su colaboración desinteresada en la investigación.

ÍNDICE GENERAL

1 CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO 13
1.1 ANTECEDENTES 13
1.2 DEFINICIÓN DEL PROBLEMA 13
1.3 OBJETIVOS 14
1.3.1 OBJETIVO GENERAL 14
1.3.2 OBJETIVOS ESPECÍFICOS 14
1.4 JUSTIFICACIÓN 15
1.5 ALCANCES 18
1.6 LIMITACIONES 18
2 CAPÍTULO II: ESTADO DEL ARTE 19
2.1 REVISIÓN DE LA LITERATURA 19
2.2 TÉCNICAS PREVIAMENTE APLICADAS 24
2.2.1 Máquinas de Vectores de Soporte (SVN) 24
2.2.2 Algoritmo K-Means 26
2.2.3 Algoritmo KNN 27
2.2.4 Árboles de Decisión 28
2.2.5 Sistemas Expertos 32
2.2.6 Redes Neuronales 37
2.3 CASOS DE ÉXITO 38
2.3.1 Reconocimiento de Patrones en imágenes citológicas cervicales en 2D para la detección temprana del cáncer cervical (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009) 38
2.3.2 Aplicación de Redes Bayesianas Dinámicas para la predicción de cáncer cervical (Onísko, Druzdzel, & Austin, 2009) 41

2.3.3 Sistema experto para diagnóstico temprano de cáncer de cuello uterino (Sanchez, 2012) 43

2.3.4 La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino (Pereira & Chamorro, 2012) 45

2.3.5 Predicción de recurrencia en pacientes con cáncer de cuello uterino utilizando MARS y Clasificación (Chang, Cheng, Lu, & Liao, 2013) 48

2.3.6 Identificación de regiones anormales en la zona cervical utilizando imágenes del examen de colposcopía (Liang, Zheng, Huang, Milledge, & Tokuta, 2013) 50

2.3.7 Prevención y detección de cáncer utilizando técnicas de Minería de Datos (Ramachandran, Girija, & Bhuvaneswari, 2014) 52

2.3.8 Redes Bayesianas para el apoyo en el diagnóstico de pacientes evaluados con el test de Papanicolaou (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014) 54

2.3.9 Detección y clasificación de cáncer cervical utilizando análisis de texturas (Soumya, Sneha, & Arunvinodh, 2016) 56

2.3.10 Predicción del cáncer de cuello uterino mediante inducción híbrida (Vidya & Nasira, 2016) 58

2.3.11 Método de identificación de cáncer cervical sobre imágenes de histología basado en las características de textura y el área de lesión (Wei, Gan, & Ji, 2017) 59

2.3.12 Aplicación de Deep Learning para la clasificación de imágenes obtenidas por colposcopía (Sato, y otros, 2018) 60

3 CAPÍTULO III: MÉTODO PROPUESTO – REDES BAYESIANAS 63

3.1 JUSTIFICACIÓN 63

3.2 METODOLOGÍA 70

3.2.1 FASE 1: SELECCIÓN DEL CONJUNTO DE DATOS 70

3.2.2 FASE 2: TRATAMIENTO DE DATOS 70

3.2.3 FASE 3: CONSTRUCCIÓN DEL MODELO PROBABILÍSTICO 71

3.3 EJEMPLO UTILIZANDO EL MÉTODO PROPUESTO 72

4 CAPÍTULO IV: ANÁLISIS, DISEÑO E IMPLEMENTACIÓN DEL SISTEMA 76
4.1 SELECCIÓN DEL CONJUNTO DE DATOS 76
4.2 TRATANMIENTO DE DATOS 76
4.3 CONSTRUCCIÓN DEL MODELO PROBABILÍSTICO 80
5 CAPÍTULO V: EXPERIMENTOS NUMÉRICOS 88
6 CAPÍTULO VI: CONCLUSIONES Y TRABAJOS FUTUROS 93
7 REFERENCIAS BIBLIOGRÁFICAS 95

ÍNDICE DE FIGURAS

Figura 1.1. Oportunidades perdidas para hacer pruebas de detección de cáncer cervical (Centros para el Control y la Prevención de Enfermedades, 2014).....15
Figura 1.2. Mortalidad a causa del cáncer de cuello uterino desde 1975 hasta el 2011 (Centros para el Control y la Prevención de Enfermedades, 2014). 16
Figura 1.3. Mortalidad estimada por cáncer de cuello uterino en el mundo (World Health Organization, 2013). 17
Figura 2.1 Representación geométrica de hiperplano generado por las SVM (Carmona Sánchez, 2014).....24
Figura 2.2 Clasificador Lineal utilizado por las SVM (Carmona Sánchez, 2014).....25
Figura 2.3 Ejemplo de clustering utilizando K-means (McCullock, 2013).....27
Figura 2.4 Ejemplo de clasificación utilizando KNN (DeWilde, 2012). 28
Figura 2.5 Ecuación para el cálculo del factor Gini (Will, 2016). 30
Figura 2.6 Diferentes árboles de decisión generados por el algoritmo RFT (Hatcher, 2014) 30
Figura 2.7 Pseudocódigo correspondiente al algoritmo ID3 (Gómez Fernández, 2013) 31
Figura 2.8 Ejemplo de cálculo de entropía (Sayad, 2012).....31
Figura 2.9 Teorema de Bayes (Juárez Ibujes, 2016). 33
Figura 2.10 Ejemplo de probabilidades a priori y posteriori de una Red Bayesiana (Carvalho, 2015).....34
Figura 2.11 Ejemplo de reglas para sacar dinero de un cajero automático (Sancho Caparrini, 2017).....35
Figura 2.12 Modus Ponens y Modus Tollens (Sancho Caparrini, 2017). 36
Figura 2.13 Estructura básica de una neurona (Introducción a las redes neuronales artificiales, s.f.).....38
Figura 2.14. Función ondícula asociada a la extracción de características de una imagen. (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009). 39
Figura 2.15. Aproximación de coeficientes de la función ondícula. (a) célula normal, (b) célula cancerígena (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009).....40
Figura 2.16. Representación gráfica del PCCSM con 19 variables (Onísko, Druzdzel, & Austin, 2009). 42

Figura 2.17. Resultados de riesgo de un paciente evaluado durante 15 años (Onísko, Druzdzel, & Austin, 2009). 43
Figura 2.18. Árbol de dominio utilizado en las reglas de inferencia (Sanchez, 2012). 44
Figura 2.19. Resultados de clasificación en Weka (Pereira & Chamorro, 2012). 47
Figura 2.20. Resultados de asociación en Weka (Pereira & Chamorro, 2012). 47
Figura 2.21. Resultados obtenidos en cada una de las 10 etapas de pruebas independientes (Chang, Cheng, Lu, & Liao, 2013). 49
Figura 2.22. Algoritmo propuesto para remover la reflexión especular (Liang, Zheng, Huang, Milledge, & Tokuta, 2013). 51
Figura 2.23. Segmentación de regiones Rojo (regiones con sangre), Verde (regiones oscuras), Azul (otras regiones) (Liang, Zheng, Huang, Milledge, & Tokuta, 2013). 51
Figura 2.24. Algoritmo propuesto: Árboles de decisión, Clustering y K-Means (Ramachandran, Girija, & Bhuvaneswari, 2014). 53
Figura 2.25. Árbol de decisión generado por división binaria recursiva (Ramachandran, Girija, & Bhuvaneswari, 2014). 53
Figura 2.26. Representación gráfica de la Red Bayesiana utilizada por Bountris et al (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014). 55
Figura 2.27. Probabilidades asociadas a algunas evidencias (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014). 55
Figura 2.28. Esquema del sistema propuesto por Soumya et al (Soumya, Sneha, & Arunvinodh, 2016). 56
Figura 2.29. Comparación del nuevo modelo propuesto utilizando procesamiento de imágenes versus el modelo actual utilizando solo factores clínicos (Soumya, Sneha, & Arunvinodh, 2016). 57
Figura 2.30. Resultados obtenidos luego de la aplicación de CART, RFT y RFT + K-Means (Vidya & Nasira, 2016). 59
Figura 2.31. Propuesta desarrollada por Wei, Gan y Ji (Wei, Gan, & Ji, 2017). 60
Figura 2.32. Arquitectura de la Red Neuronal utilizada para evaluar las imágenes de la zona cervical (Sato, y otros, 2018). 61
Figura 3.1. Resultados obtenidos por las SVM y las Redes Bayesianas (Rubio, Martínez-Gómez, Flores, & Puerta, 2016). 67

Figura 3.2. Evidencia recolectada entre el 2005 y 20015 sobre la aplicación de Redes Bayesianas (Langarizadeh & Moghbeli, 2016). 69
Figura 3.3. Representación gráfica de la metodología (Elaboración Propia) 70
Figura 3.4. Conjunto de datos a utilizar en el ejemplo (Puga, 2012). 73
Figura 3.5. Estructura de la Red Bayesiana de ejemplo (Puga, 2012). 74
Figura 3.6. Estructura de las Red Bayesiana con los posibles valores de cada variable (Puga, 2012). 74
Figura 3.7. Fórmula utilizada para el cálculo de las probabilidades asociadas (Puga, 2012) 75
Figura 3.8. Tablas de probabilidad asociadas a la RB. Enfermedad (arriba), Problemas Respiratorios (izquierda), Dolor de cabeza (derecha). 75
Figura 4.1.Topología de la Red Bayesiana en construcción (Elaboración Propia) 82
Figura 4.2. Consola de Netica con las iteraciones del algoritmo EM (Elaboración Propia) 82
Figura 4.3. Estado de los nodos de la Red Bayesiana luego del Aprendizaje Paramétrico (Elaboración Propia). 83
Figura 4.4. Tabla de probabilidad para la variable #Parejas Sexuales (Elaboración Propia) 85
Figura 4.5. Tabla de probabilidad para la variable #Embarazos (Elaboración Propia) 85
Figura 4.6. Tabla de probabilidad de la variable #ETS (Elaboración Propia). 86
Figura 4.7. Tabla de probabilidad de la variable #Años utilizando DIU (Elaboración Propia). 86
Figura 4.8. Tabla de probabilidad de la variable #Años utilizando anticonceptivos hormonales (Elaboración Propia). 87
Figura 4.9. Extracto de la tabla de probabilidad para la variable target (Elaboración Propia). 87
Figura 5.1. 30 casos de prueba utilizados para la validación y resultados (Elaboración propia). 88
Figura 5.2. Ecuación para envía mensajes ascendentes (Elaboración Propia). 89
Figura 5.3. Ecuación para enviar mensajes descendentes (Elaboración Propia). 89
Figura 5.4. Ingresando los registros a ser utilizados en la validación (Elaboración Propia). 89

ÍNDICE DE TABLAS

Tabla 2.1 Ventajas y Desventajas de la SVM (Elaboración Propia). 26
Tabla 2.2 Ventajas y Desventajas de las Redes Bayesianas (Elaboración Propia). 34
Tabla 2.3 Ventajas y Desventajas de los Sistemas Expertos basados en reglas (Elaboración Propia) 37
Tabla 3.1 Valores y puntajes de los criterios de comparación (Elaboración propia). 65
Tabla 3.2 Benchmarking de las diferentes técnicas analizadas (Elaboración Propia)..... 66
Tabla 4.1. Variables utilizadas en el trabajo de investigación (Elaboración Propia) 78
Tabla 4.2. Variables utilizadas y sus posibles valores luego de la discretización. (Elaboración Propia) 80
Tabla 5.1. Matriz de confusión de la predicción (Elaboración Propia)..... 90
Tabla 5.2. Características obtenidas a partir de la Matriz de Confusión (Elaboración propia). 92

RESUMEN

El cáncer de cuello uterino una de las principales causas de muerte por cáncer en las mujeres. Una gran variedad de técnicas utilizadas en la Inteligencia Artificial (IA) como las Redes Neuronales, las Máquinas de Vectores de Soporte (SVM), los Árboles de Decisión y otros; han abordado el problema de la predicción de esta enfermedad. El siguiente trabajo de investigación realiza la predicción de riesgo de cáncer de cuello uterino usando un modelo probabilístico basado en Redes Bayesianas; donde de un total de 322 registros se pudo obtener 15 atributos o características diferentes que correspondan a la información de una paciente. Las pruebas fueron realizadas utilizando el 40% de los datos. Los resultados le otorgan al trabajo desarrollado una tasa de éxito del 96%, además, sugieren que las Redes Bayesianas alcanzan un alto rendimiento, así como también ofrecen transparencia durante el proceso de inferencia, algo que no sucede con muchas otras técnicas, y que son ideales para afrontar problemas de predicción.

Palabras Claves: Modelos Predictivos, Método de Bayes, Algoritmos de Predicción, Computación Probabilística

ABSTRACT

Cervical cancer is one of the main causes of death due to cancer in women. A large number of techniques from the Artificial Intelligence (AI) such as Neuronal Networks, Support Vector Machines (SVM), Decision Trees and others; have been used to deal with the problem of predicting this disease. The following research assess the cervical cancer risk prediction, by implementing a probabilistic model based on Bayesian Networks and using 322 instances where we could retrieve 15 different features that are known information from each patient. The tests were made using the 40% of the whole dataset. The results show that this work has raised a 96% of success rate, in addition to this, the results suggest that Bayesian Networks are able to reach a high performance and provide transparency during the inference process at the same time, something that does not happen in many other techniques, and that they are really efficient to face this sort of prediction problems.

Keywords: Predictive models, Bayes methods, Prediction algorithms, Probabilistic Computing

INTRODUCCIÓN

El trabajo realizado durante las últimas décadas por investigadores procedentes de varios campos de la Inteligencia Artificial muestra cómo muchos de los problemas que antes parecían imposibles, o intratables, pueden hoy ser formulados y resueltos por máquinas (Castillo, Gutiérrez, & Hadi, 1997).

El desarrollo de sistemas de diagnóstico basados en técnicas bayesianas comenzó en los años 60. Tal y como señala F. J. Díez (Díez, 1998), estos primeros sistemas utilizaban el método probabilístico clásico, que, a pesar de presentar algunas deficiencias, abría paso a futuras implementaciones, técnicas y mejoras basadas en probabilidades como lo son hoy en día las Redes Bayesianas, un modelo gráfico probabilista (MGP) que sería visto por primera vez en los años 80 y a partir de ahí pasaría a ser la base de muchos sistemas expertos probabilísticos. Según F. J. Díez (Díez, 1998), los MGP encajan perfectamente con la medicina, dado que esta última posee dos propiedades importantes: el conocimiento causal y las fuentes de incertidumbre. Precisamente el presente trabajo de investigación aborda la predicción de riesgo de cáncer cervical con la ayuda de un MGP, como lo son las redes bayesianas.

El diagnóstico para el cáncer de cuello uterino requiere de una biopsia que debe ser tomada de una lesión cervical visible (INEM, 2013). Existen diferentes procedimientos para llevarlo a cabo siendo el Papanicolaou uno de los más conocidos y que la Organización Mundial de la Salud (OMS) recomienda realizar a las mujeres mayores de treinta años con cierta frecuencia (Organización Mundial de la Salud, 2019). En los países desarrollados, se han puesto en marcha programas que han logrado prevenir hasta el 80% de los casos de cáncer de cuello uterino, sin embargo, esto es algo que aún no se logra en los países en desarrollo. De esta forma es que se constituye el problema con el que la presente investigación pretende lidiar.

1 CAPÍTULO I: PLANTEAMIENTO METODOLÓGICO

1.1 ANTECEDENTES

El cáncer de cuello uterino, también conocido como cáncer cervical, es una clase común de cáncer en la mujer, es una enfermedad en la cual se detectan células cancerosas (malignas) en los tejidos del cuello uterino. Este tipo de cáncer suele crecer lentamente por un período de tiempo. Antes de que se encuentren células cancerosas en el cuello uterino, sus tejidos experimentan cambios y empiezan a aparecer células anormales. Posteriormente, las células cancerosas comienzan a crecer y se dispersan con mayor profundidad en el cuello uterino y en las áreas circundantes.

Internacionalmente, las tres primeras causas de muerte por cáncer en mujeres corresponden en orden descendente a cáncer de mama, cáncer de pulmón y cáncer cervical con tasas estandarizadas por edad entre 20 y 50 años.

Mundialmente, las más afectadas por esta patología son las mujeres de escasos recursos, que tienen menos acceso a la detección precoz. Los datos que dispone la OMS indican que las tasas de cáncer de cuello uterino son mayores en los países del Tercer Mundo, especialmente en América Latina durante el 2005; sin embargo, en los países subdesarrollados, el cáncer de cuello uterino ocupa el segundo lugar entre las causas de muerte por cáncer en la mujer.

La detección primaria de cáncer de cuello uterino se hace por medio de exámenes clínicos como el Papanicolaou, examen que ayuda a detectar células anormales en el revestimiento del cuello uterino antes de que puedan convertirse en células pre cancerosas o cancerígenas.

1.2 DEFINICIÓN DEL PROBLEMA

Los resultados estadísticos presentados por la OMS calculan que en el 2012 hubo 530,000 nuevos casos de cáncer de cuello uterino, los cuales representaron el 7.5% de la mortalidad femenina, siendo la falta de prevención la principal causa.

En el 2013, según la OMS, los nuevos adelantos tecnológicos ofrecen la posibilidad de enfrentar el cáncer de cuello uterino de una manera mucho más integral con la finalidad de prever un futuro más saludable para los niñas y mujeres. Sin embargo, en los países menos

desarrollados la realidad es que se carecen de sistema de salud eficaces y de recursos financieros suficientes en comparación con los países desarrollados. Se han identificado los siguientes problemas comunes durante el diagnóstico del cáncer de cuello uterino:

- En los países subdesarrollados no se cuenta con el personal capacitado para realizar el análisis respectivo y entregar el diagnóstico al paciente. Esto provoca retrasos en la entrega de resultados y genera pérdida de interés en los pacientes.
- Cuando la paciente es diagnosticada como negativo en las pruebas de cáncer de cuello uterino, suele olvidarse la importancia de realizar el examen cada cierto tiempo, lo que puede terminar causando que la enfermedad se presente después y sea detectada de forma tardía.
- Un alto porcentaje de mujeres no acude al ginecólogo por vergüenza a realizarse los exámenes y en muchos casos por los altos costos de atención.

1.3 OBJETIVOS

1.3.1 OBJETIVO GENERAL

Desarrollar un Sistema Experto Probabilístico basado en Redes Bayesianas para la predicción de riesgo de cáncer cervical, el cual permitirá conocer el riesgo de cada paciente de desarrollar esta enfermedad permitiendo a los doctores agilizar el proceso de diagnóstico y contribuir a realizarles un mejor control.

1.3.2 OBJETIVOS ESPECÍFICOS

- Estudiar y revisar las diferentes técnicas y/o modelos existentes enmarcados en el campo de la Inteligencia Artificial con los cuáles se podría ofrecer una solución al caso en estudio.
- Desarrollar un sistema confiable con una precisión no menor del 80%.
- Evaluar la efectividad del sistema utilizando datos de pacientes reales que alguna vez participaron en el proceso de diagnóstico de cáncer cervical.
- Ofrecer visibilidad y transparencia a los usuarios sobre el proceso de evaluación de riesgo, de tal forma que la herramienta no sea una caja negra para ellos, sino por el contrario, que pueda ser utilizado como punta de partida para otros estudios.

1.4 JUSTIFICACIÓN

Como bien se ha mencionado ya, hoy en día se cuentan con muchas pruebas para detectar tempranamente el cáncer de cuello uterino que podrían salvar la vida de muchas mujeres. Según los Centros para el control y Prevención de Enfermedades (CDC), en el 2014 existían alrededor de 8 millones de mujeres entre 21 y 65 años que no se habían realizado alguna prueba de detección de cáncer de cuello uterino durante los últimos 5 años (Centros para el Control y la Prevención de Enfermedades, 2014). A continuación, la Figura 1.1 muestra la cantidad de mujeres que acuden al médico regularmente, y sin embargo, no se realizan ninguna prueba para la detección temprana de cáncer de cuello uterino.

Figura 1.1. Oportunidades perdidas para hacer pruebas de detección de cáncer cervical (Centros para el Control y la Prevención de Enfermedades, 2014).

La subdirectora principal de CDC, Ileana Arias, afirma que ninguna mujer debería morir por causa del cáncer de cuello uterino, y que cada visita médica es una nueva oportunidad de prevenir esta enfermedad. La Figura 1.2 muestra gráficamente cómo la mortalidad disminuyó notablemente a lo largo de los años hasta el 2007, sin embargo, a partir de ese año en adelante, el progreso parece haberse quedado estancado.

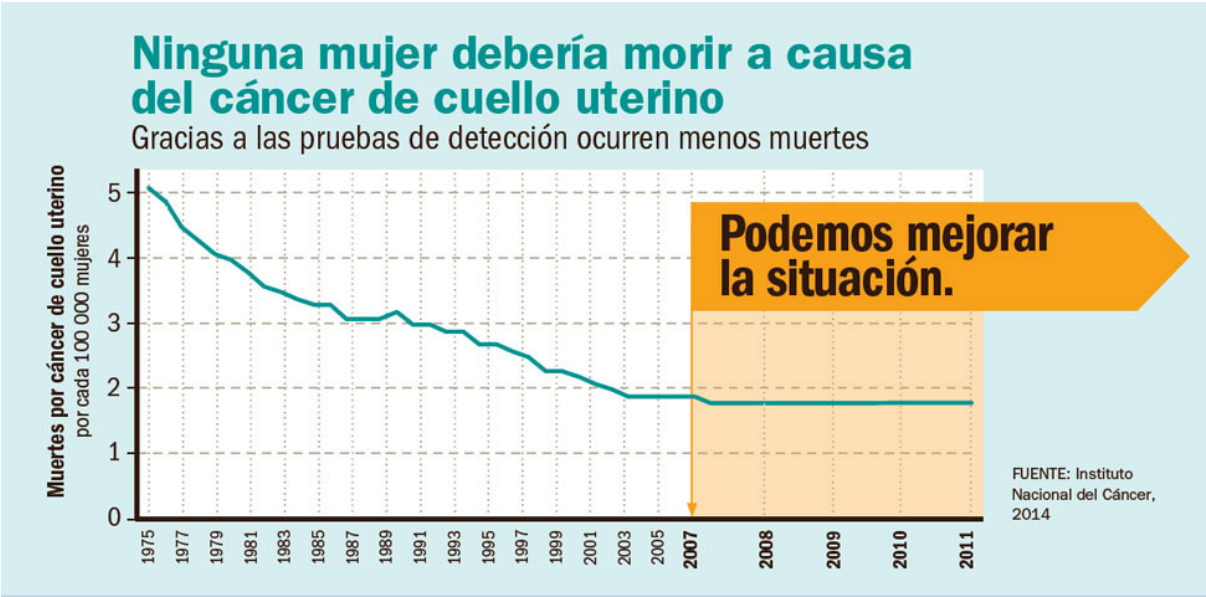


Figura 1.2. Mortalidad a causa del cáncer de cuello uterino desde 1975 hasta el 2011 (Centros para el Control y la Prevención de Enfermedades, 2014).

En Estados Unidos, un grupo de investigadores realizaron una revisión de la cantidad de casos de cáncer de cuello uterino que tuvieron lugar entre el 2007 y 2012. Los hallazgos claves fueron descritos de la siguiente forma:

- En el 2012, el 11.4% de las mujeres reportó no haberse hecho pruebas de detección de cáncer de cuello uterino en los últimos cinco años. El porcentaje fue mayor en las mujeres sin seguro médico (23.1%) y en aquellos sin un proveedor de atención médica habitual (25.5%).
- El porcentaje de mujeres que no se habían hecho pruebas de detección según lo recomendado fue más alto entre las mujeres mayores (12.6%), las asiáticas o isleñas del Pacífico (19.7%) y las indoamericanas o nativas de Alaska (16.5%).
- Entre el año 2007 y 2011, la tasa de incidencia de cáncer de cuello uterino disminuyó en 1.9% por año mientras que la tasa de mortalidad permaneció estable.
- La zona sur tuvo la tasa más alta de cáncer de cuello uterino (8.5 por cada 100, 000), la tasa de mortalidad más alta (2.7 por cada 100, 000) y el mayor porcentaje de mujeres que no se habían hecho pruebas de detección en los últimos cinco años (12.3%).

Incluso con las mejoras hechas en los métodos de prevención y detección temprana, la mayoría de casos de cáncer de cuello uterino se da en mujeres que no se encuentran al día con sus pruebas, lo cual se debe a factores económicos en muchos casos.

Según el CNEGSR, un estimado de 528 mil nuevos casos son diagnosticados anualmente, 85% de los cuales se registran en países en vía de desarrollo y 266 mil defunciones anuales, 87% de las cuales ocurren en países subdesarrollados. Cabe resaltar que la tendencia de la mortalidad es descendente respecto a años anteriores debido a una menor incidencia de la enfermedad producto de la mejora en las técnicas de detección y prevención, lo cual constituye un indicador de desigualdad, pues la mortalidad tiende a concentrarse en las regiones más desfavorecidas.

La OMS en el 2013 (World Health Organization, 2013) informó que el cáncer de cuello uterino es el cáncer más frecuente en mujeres en 45 países del mundo y mata a más mujeres que cualquier otra de cáncer en 55 países, entre ellos muchos países del África, otros de Asia y algunos centroamericanos y sudamericanos (Ver Figura 1.3).

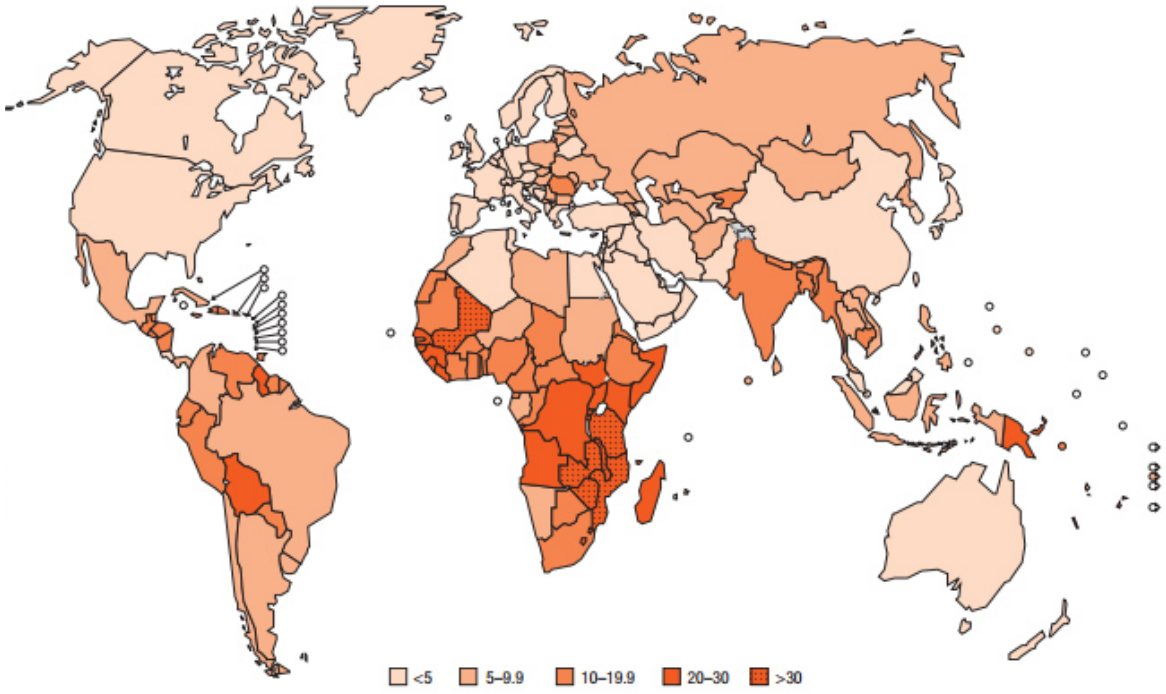


Figura 1.3. Mortalidad estimada por cáncer de cuello uterino en el mundo (World Health Organization, 2013).

Basado en los datos mostrados se puede inferir que a pesar de que la tasa de mortalidad se ha disminuido considerablemente, aún se puede trabajando para reducirla aún más, sobre todo en las zonas donde el nivel socioeconómico no es muy alto, y si carecen de instrumentos y especialistas que para abstener la atención eficiente de las pacientes. Es por ello que en el presente trabajo de investigación se desarrollará un sistema capaz de predecir el riesgo de desarrollar cáncer de cuello.

1.5 ALCANCES

- La implementación, el modelo y documentación sobre la implementación serán liberados mediante un repositorio en GitHub aceptando sugerencias, mejoras y contribuciones.
- El modelo probabilístico será completamente transparente para el usuario final permitiéndole analizar la interacción entre cada una de las variables que influyen en el resultado.

1.6 LIMITACIONES

- El modelo de la Red Bayesiana utiliza a lo mucho 16 variables, por restricciones de la versión Free Limited Edition del software Netica.
- La estructura de la Red Bayesiana está pensando exclusivamente para modelar el conocimiento del dominio correspondiente al cáncer de cuello uterino, pero no se descarta que la estructura pueda ser utilizada como punto de inicio para plantear el análisis de otros dominios distintos.

2 CAPÍTULO II: ESTADO DEL ARTE

En el presente capítulo se hará una recopilación de diversos casos de éxito aplicando diferentes métodos para solucionar un problema en común: la predicción y/o detección de cáncer cervical o cáncer de cuello uterino. En el subcapítulo 2.1 se llevará a cabo una revisión general de todos los casos de éxito ordenados cronológicamente, mientras que en el subcapítulo siguiente se describirán los diferentes métodos que han sido utilizados para afrontar el problema en estudio. Finalmente, en el subcapítulo 2.3, se tratará con detalle cada una de las diferentes fuentes científicas que sirvieron de base para alimentar el Estado del Arte de la presente investigación.

2.1 REVISIÓN DE LA LITERATURA

En el año 2009 se presentaron dos casos de estudio relacionados al cáncer de cuello uterino, uno de ellos en Indonesia, donde el Departamento de Ingeniería Biomédica de la Universidad Swiss German (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009) presentó un sistema para la detección temprana de cáncer de cuello uterino utilizando reconocimiento de patrones sobre imágenes citológicas de la zona cervical en 2D, con el fin de lidiar con el problema de inconsistencia e imprecisión del diagnóstico del test del Papanicolaou. Por otro lado, Onísko, Druzdzel y Marshall implementaron un modelo basado en Redes Bayesianas Dinámicas al cual dieron el nombre de PCCSM (Pittsburgh Cervical Cancer Screening Model), el cual permite evaluar el riesgo de cáncer cervical invasivo en el transcurso del tiempo, con la finalidad de obtener una serie de resultados por paciente en diferentes etapas cronológicas, lo cual ayudaría a los médicos a realizar un seguimiento individual a cada uno de ellos para tomar decisiones más certeras al momento de programar las evaluaciones médicas respectivas (Onísko, Druzdzel, & Austin, 2009).

En el año 2012, nuevamente se presentaron dos casos de estudio muy interesantes. El primero fue desarrollado en Perú, por la alumna de pregrado Barturen Sánchez, de la escuela de Ingeniería de Sistemas de la Universidad Católica Santo Toribio de Mogrovejo. El trabajo de investigación consistió en la implementación de un sistema experto basado en reglas de inferencia para el diagnóstico temprano de cáncer de cuello uterino. A pesar de contar con una cantidad de datos relativamente pequeña (113 historias clínicas), el sistema experto logró

alcanzar un grado de confianza del 97%, un número bastante alto y muy alentador (Sanchez, 2012). Mientras tanto, Ricardo Timarán Pereira y María Clara Yépez Chamorro desarrollarlo en conjunto con la Universidad de Nariño, Colombia, un estudio cuya finalidad fue descubrir patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. Para llevar el proyecto a cabo se vieron en la necesidad de utilizar diferentes técnicas de clasificación y asociación basadas en árboles de decisión, siendo capaces de generar modelos bastante consistentes con la realidad observada. El estudio llegó a la conclusión que el 95.4% de mujeres muertas a causa de cáncer de cuello uterino cumplían con tres características importantes: no ser madre cabeza de familia, no estar clasificada en el SISBEN (Sistema de Selección de Beneficiarios para Programas Sociales) y no pertenecer a algún régimen de salud (Pereira & Chamorro, 2012).

Al año siguiente, 2013, dos nuevos casos de estudio se dieron a conocer. La Revista Internacional de Machine Learning y Computación (IJMLC por sus siglas en inglés) publicó un estudio enfocado en el problema de la recurrencia del cáncer de cuello uterino en pacientes que alguna vez fueron diagnosticados con esta enfermedad (Chang, Cheng, Lu, & Liao, 2013). Haciendo uso de un conjunto de datos correspondiente a las historias clínicas de 168 pacientes fueron capaces de desarrollar dos modelos diferentes, ambos con técnicas diferentes. Uno de ellos utilizó el algoritmo C5.0 mientras el otro utilizó el algoritmo de clasificación MARS. Los resultados demostraron que la efectividad de los árboles de decisión + C5.0 alcanzan un 96% de nivel de confianza superando al modelo de MARS que alcanzó un 86%. De la misma manera, en la Conferencia Internacional de Computación Gráfica, Visualización y Visión Computacional se presentó un estudio acerca de la identificación de regiones anormales en la zona cervical haciendo uso de imágenes extraídas del examen de colposcopía (Liang, Zheng, Huang, Milledge, & Tokuta, 2013). Las pruebas realizadas sobre un conjunto de 48 imágenes pertenecientes a 12 pacientes con diferentes diagnósticos respecto al cáncer cervical, demostraron que, las Máquinas de Vectores de Soporte (SVM por sus siglas en inglés) alcanzan una tasa de éxito que asciende al 94.6% haciendo uso de una función Kernel lineal, que a pesar de ser la más simple, se comportó muchísimo mejor que funciones más complejas como la función de base radial o la función polinomial.

El año 2014, la Revista Internacional de Aplicaciones Computacionales (IJCA por sus siglas en inglés) presentó un estudio realizado en Chennai, India, en el cual se logró implementar un modelo de detección temprana de cáncer haciendo uso de técnicas de Minería de Datos, específicamente de los famosos Árboles de decisión y de la técnica de Clustering: K-Means. Utilizando 746 casos de prueba, el 99% fue clasificado correctamente (Ramachandran, Girija, & Bhuvaneswari, 2014). En el mismo año, en la conferencia MobiHealth, se dio a conocer un sistema basado en Redes Bayesianas cuyo el objetivo fue el de predecir el diagnóstico de pacientes que fueron sometidos a pruebas como el Papanicolaou y obtuvieron un resultado dudoso o ambiguo. Este sistema resultó bastante prometedor alcanzando un porcentaje de 94.9% en términos de Valor Predictivo Positivo y 95.5% en términos de Valor Predictivo Negativo (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014).

Dos años más tarde, en el 2016, en Calicut, India, el Departamento de Ciencias de la Computación e Ingeniería de la Universidad Akkikavu presentó una investigación basada en la detección y clasificación de cáncer cervical analizando la textura de imágenes extraídas del examen de Resonancia Magnética (MRI por sus siglas en inglés). Haciendo uso de modelos de clasificación basados en SVM no lineales y de los atributos recolectados de una serie de transformaciones y algoritmos aplicadas a las imágenes, como la transformación de Contourlet, o las características de Gabor, se concluyó que las imágenes obtenidas por MRI de tipo Sagital Ponderadas T2 favorecen al buen desempeño del modelo alcanzando una precisión del 83%, ligeramente superior a resultados obtenidos con imágenes Axiales Ponderadas T1 y T2 que obtuvieron 81%y 82% respectivamente (Soumya, Sneha, & Arunvinodh, 2016). Por otro lado, también en la India, la Universidad de Manonmaniam Sundaranar realizó una comparación entre distintos métodos aplicados para la predicción del cáncer de cuello uterino utilizando datos genéticos (Vidya & Nasira, 2016). Un total de 100 registros fueron utilizados en el proyecto, de los cuáles 60 correspondieron al conjunto de datos de entrenamiento y 40 al conjunto de pruebas. Se pusieron a prueba 3 técnicas diferentes: el algoritmo de CART, con el cual se alcanzó un 83.7% de tasa de éxito; el algoritmo RFT, cuya tasa de éxito asciende a 93.54%; y finalmente un modelo que combina dos algoritmos: el RFT y el K-Means. Este último modelo obtuvo resultados muchísimo mejores que en los experimentos anteriores, logrando alcanzar una tasa de éxito de 96.77%.

En el año siguiente, 2017, la Universidad Politécnica de Anhui, China, publicó un artículo describiendo los resultados de una investigación acerca de la identificación de cáncer cervical utilizando imágenes de la zona cervical, principalmente basándose en la textura y en las áreas que presentan lesiones (Wei, Gan, & Ji, 2017). El método utilizado combina diferentes técnicas enmarcadas en el campo de la Inteligencia Artificial, tales como el K-Means que fue utilizado para la segmentación y agrupación de la zona cervical logrando distinguir las áreas afectadas, y, por último, se utilizaron SVN para reconocer si en efecto, la zona analizadas se trata de cáncer cervical o no. Los resultados experimentales muestran que la precisión alcanzada por el método propuesto asciende a un 90%, el cual, si bien es una tasa bastante alta, los autores mencionan que sería ideal comprobarlo con un número más alto de datos de prueba, dado que para los resultados expuestos únicamente se utilizaron 20 datos prueba.

En el 2018, en Japón, el Departamento de Ginecología del Centro de Cáncer de Saitama presentó un proyecto de investigación que pretende demostrar el uso de Deep Learning para abordar problemas de clasificación, en este caso en particular, la clasificación de imágenes obtenidas por el examen de colposcopía. Lamentablemente, los resultados no fueron los esperados, dado que la evaluación en términos de precisión alcanzó el 50% a diferencia de casos de estudio previamente analizados, sin embargo, los autores mencionan que es complicado comparar estos resultados con trabajos previamente implementados dado que es la primera vez que se decide aplicar Deep Learning para abordar el tema de clasificación de imágenes de colposcopía y que los resultados son lo suficientemente buenos a nivel clínico considerando la dificultad de los propios médicos para diferenciar entre diferentes displasias, incluso siendo ellos los expertos (Sato, y otros, 2018).

Un año después, en el 2019, la Revista del Instituto Nacional de Cáncer publicó un estudio que desarrolló un algoritmo capaz de automatizar el diagnóstico de displasias cervicales. Los resultados fueron bastante alentadores, siendo que, el algoritmo remitió al 91.7% de los casos de HPV positivo CIN3/AIS a un examen de colposcopía inmediata, mientras que difirió un total del 38.4% de todas las mujeres con HPV positivo a una reevaluación de un año, lo cual, comparado con un examen directo de citología cuyos resultados eran de 89.1% y 37.4% respectivamente, lo convierten en, sin duda, un algoritmo confiable para el proceso de diagnóstico o triaje como es mencionado por los autores, de una displasia cervical. Por otro

lado, en Los Ángeles, California, la Universidad del Sur de California realizó un estudio que fue publicado en la Revista Americana de Obstetricia y Ginecología el cual consiste en la comparación de dos modelos diferentes, uno basado en el modelo de riesgo proporcional de Cox, y otro basado en Deep Learning para analizar las probabilidades de supervivencia de pacientes diagnosticados con cáncer de cuello uterino. Los resultados ofrecidos sugieren que el modelo basado en Deep Learning es una herramienta confiable para predecir la supervivencia de mujeres con cáncer de cuello uterino basándose en el error absoluto obtenido por cada uno de los diferentes modelos, 43.6% con el modelo de Cox, y 30.7 con el modelo de Deep Learning, lo cual significa que su tasa de éxito ascendía al 69.3%, pero eso no es todo, sino que, al incluir más características, del total de 40 que se tenían por paciente, los resultados mejoraban alcanzando una tasa de éxito de 79.5% haciendo uso de las 40 (Matsuo, y otros, 2019). Por otro lado, en la Revista del Instituto Nacional de Cáncer se publicó un artículo que trata sobre la aplicación de Deep Learning y la evaluación automatizada de imágenes de la zona cervical para detectar cáncer de cuello uterino. Haciendo uso del enfoque de Deep Learning conocido como Faster R-CNN, se entrenó un modelo utilizando cervigramas digitalizados. El algoritmo en mención alcanzó un alto nivel de precisión, cuyo valor asciende al 91%, superando de esta manera incluso a la interpretación original de los cervigramas que contaban con un 69% o a la evaluación citología convencional, que contaba con un 71% (Hu, y otros, 2019). Asimismo, en Pakistán, 2 universidades de la ciudad de Lahore publicaron un artículo en la Revista Internacional de Ciencias de la Computación Avanzada y Aplicaciones comparando diferentes métodos para predecir el cáncer de cuello uterino haciendo uso de Minería de Datos. Se utilizaron tres técnicas diferentes para compararlas entre sí, las cuales fueron: Árboles de Decisión, Bosque de Decisión y Jungla de Decisión. Los resultados fueron que, en todos los experimentos, los Árboles de Decisión superaron a las otras 2 técnicas mencionadas. Se realizaron 4 experimentos tomando 4 variables diferentes del conjunto de datos como objetivo, cuando la variable biopsia era la variable objetivo, los Árboles de decisión alcanzaron un 97.4% de precisión, cuando la variable citología fue considerada como la variable objetiva, la precisión alcanzó el 95.9%, para la variable Test de Schiller como objetivo, se obtuvo 94.3% y finalmente para la variable Test de Hinselmann, la precisión ascendió a 97.8%, todos

resultados bastante alentadores a favor de los Árboles de Decisión (Mahboob, Milhan, Iqbal, Wahab, & Mushtaq, 2019).

2.2 TÉCNICAS PREVIAMENTE APLICADAS

2.2.1 Máquinas de Vectores de Soporte (SVN)

Las Máquinas de Vectores de Soporte (SVN por sus siglas en inglés), son un conjunto de algoritmos de aprendizaje supervisado que sirven para resolver problemas de clasificación y regresión. Son consideradas como clasificadores lineales dado que toma una decisión basada en un vector o combinación lineal de características (Carmona Sánchez, 2014).

La idea básica detrás de las SVM es ubicar un hiperplano que logre separar las instancias más cercanas de cada clase, con un margen máximo a cada lado del hiperplano, tal y como se puede apreciar en la Figura 2.1.

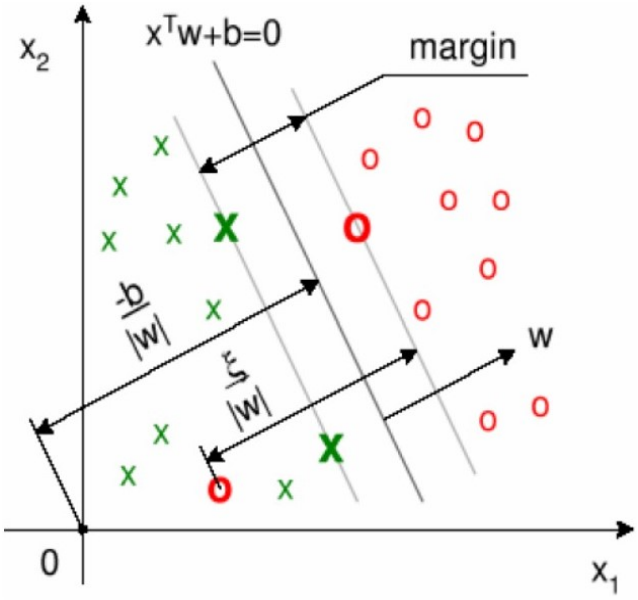


Figura 2.1 Representación geométrica de hiperplano generado por las SVM (Carmona Sánchez, 2014).

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento, las SVM se enfocan en minimizar lo que se denomina como riesgo estructural.

El clasificador lineal utilizado por las SVM permite ubicar los valores de entrada, con la ayuda de un umbral, en alguna de las clases. La Figura 2.2 corresponde a la definición del clasificador lineal.

$$y(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b$$

Figura 2.2 Clasificador Lineal utilizado por las SVM (Carmona Sánchez, 2014)

De la ecuación correspondiente al clasificador lineal:

- α_i: Multiplicador de Lagrange
- y_i: Etiqueta
- N: Número de elementos del conjunto de entrenamiento
- k (x, x_i): Función Kernel
- b: Bias

Las SVM admiten diferentes tipos de funciones Kernel, una de ellas es la función lineal, la cual es, además, la más conocida y sencilla de implementar, sin embargo, existen más alternativas como: el Kernel Polinomial, Gaussiano, Perceptrón y Sigmoidal. A continuación, la Tabla 2.1 muestra una comparativa entre las ventajas y desventajas de las SVM.

Ventajas	Desventajas
Está garantizada la ubicación de un mínimo global, a diferencia de otros métodos en los que se suelen obtener mínimos locales. Es útil para resolver tanto problemas de clases linealmente separables como no linealmente separables. Existen una innumerable cantidad de implementaciones diferentes, se puede	En el Procesamiento de Lenguaje Natural, son ampliamente superadas por otras técnicas, como las representaciones estructuradas, que se acomodan más al problema. Al igual que muchas otras técnicas y/o algoritmos, los diferentes Kernels que existen para las SVM son altamente sensibles al problema de over fitting.

utilizar la que mejor se acomode a la situación o problema.	

Tabla 2.1 Ventajas y Desventajas de la SVM (Elaboración Propia).

2.2.2 Algoritmo K-Means

K-Means es uno de los algoritmos de aprendizaje no supervisado más simples que existen para resolver problemas de clustering o agrupación.

El objetivo de este algoritmo es identificar una cantidad K de grupos en un determinado conjunto de datos. La forma en la que este algoritmo trabaja es iterativa, con la finalidad de asignar cada uno de los elementos del conjunto de datos a uno de los K grupos. Cada uno de los elementos están representados por puntos, los cuáles se agrupan en función de la similitud que hay entre ellos (Trevino, 2016). Una vez que el algoritmo K-Means acaba su ejecución, nos ofrece los siguientes resultados:

- Los centroides de los K clusters, los cuáles pueden utilizarse para etiquetar nuevos datos.
- Etiquetas para los datos de entrenamiento (cada punto de datos se asigna a un solo grupo).

En lugar de definir grupos antes de examinar el conjunto de datos, K-Means nos permite automatizar el proceso de agrupamiento. Cada centroide de un clúster, es una colección de valores que definen los grupos resultantes. Analizar las características de cada centroide puede ser de mucha utilidad para interpretar cualitativamente qué clase de grupo representa cada uno de los clusters (Trevino, 2016). La Figura 2.3 muestra con puntos azules los centroides asociados a cada uno de los clusters, y en cada uno de ellos se puede observar un conjunto de puntos negros, que corresponden a los elementos con características similares agrupados dentro de cada cluster.

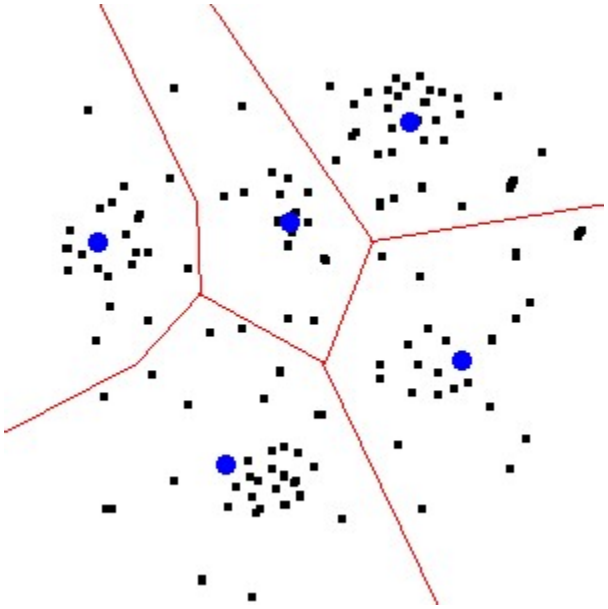


Figura 2.3 Ejemplo de clustering utilizando K-means (McCullock, 2013).

2.2.3 Algoritmo KNN

El algoritmo de los K vecinos más cercanos (KNN por sus siglas en inglés), corresponde a la clase de algoritmos de aprendizaje no paramétrico. Esto significa que el KNN no realiza ninguna suposición sobre la distribución de los datos. La utilidad de esta característica radica en la similitud con el mundo real, puesto que la mayoría de datos prácticos, no obedece a las suposiciones teóricas típicas como mezclas gaussianas, conjuntos linealmente separables, etcétera (DeWilde, 2012).

KNN también es considerado un algoritmo de aprendizaje perezoso, ya que no utiliza los datos de entrenamiento para hacer alguna generalización, en otras palabras, no existe una fase de entrenamiento explícita, o, en algunos casos es mínima. La falta de generalización significa que el KNN requiere de utilizar los datos de entrenamiento durante la fase de prueba.

La mayoría de los algoritmos perezosos, en especial el KNN toman decisiones basadas en todo el conjunto de datos de entrenamiento, o, un subconjunto de ellos. En la Figura 2.4 se puede observar los resultados de clasificar una estrella de color rojo para diferentes valores de K. Los círculos amarillos y morados, correspondientes a la clase A y B respectivamente, pertenecen al conjunto de entrenamiento. Nótese que para un valor de K = 3, la circunferencia

pequeña ubica a los 3 vecinos más cercanos, dado que en esa pequeña área prevalecen los puntos de la clase B, la estrella roja es clasificada como un elemento de la clase B. Por otro lado, para un valor de K = 6, el área de la circunferencia de expande y esta vez es la clase A la que prevalece, por lo tanto, el resultado de clasificación de la estrella roja esta vez sería de clase A (DeWilde, 2012).

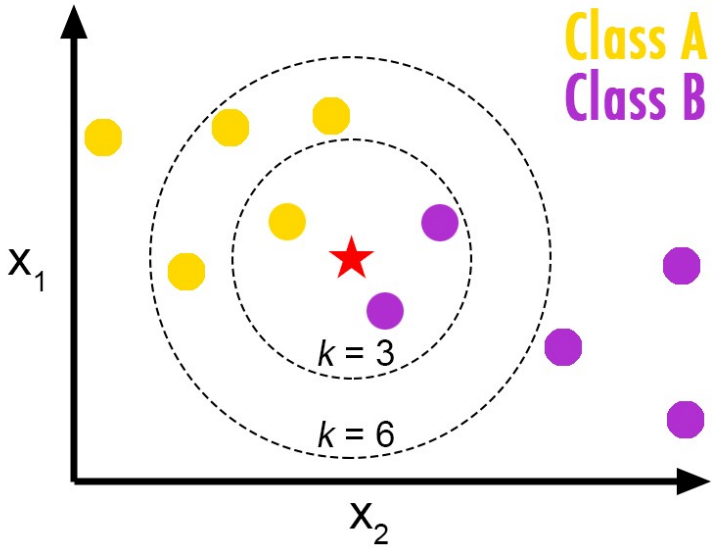


Figura 2.4 Ejemplo de clasificación utilizando KNN (DeWilde, 2012).

2.2.4 Árboles de Decisión

Los árboles de decisión son herramientas utilizadas para resolver, principalmente, problemas de clasificación. Son bastante conocidos gracias a que permiten interpretar resultados con facilidad.

Según Sancho Caparrini (2018), este tipo de estructuras ofrecen las siguientes funciones:

- **Segmentación:** Establecer cuáles son los grupos importantes para clasificar un elemento.
- **Clasificación:** Asignar un elemento a uno de los grupos (clases) existentes en un conjunto de datos.
- **Predicción:** Establecer reglas que permiten predecir ciertos eventos.
- **Reducción de la dimensión de datos:** Identificar cuáles son los datos importantes para crear modelos de determinado fenómeno.

- **Identificación-interrelación:** Identificar qué variables y relaciones son importantes para determinados grupos.
- **Recodificación:** Discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante.

Existen diferentes algoritmos utilizados para generar árboles de decisión a partir de un conjunto de datos. A continuación, se describirán los algoritmos que han sido utilizados en los trabajos previos mencionados en el presente capítulo.

2.2.4.1 Árbol de Clasificación y Regresión

El algoritmo de Árbol de Clasificación y Regresión (CART, por sus siglas en inglés) permite modelar un árbol de decisión a partir de un conjunto de datos utilizados como entrenamiento. El aprendizaje es de tipo Greedy, y se utiliza la técnica de Recursive Binary Splitting (RBS) para la generación de cada uno de los nodos (QuarkGluon Ltd, 2018).

Es importante resaltar que una de los puntos más importantes a tener en cuenta al aplicar este algoritmo es el de definir un buen criterio para detener el aprendizaje, ya que esto determinará la efectividad del árbol de decisión generado. Igualmente se debe considerar el uso de algunas técnicas de poda de árboles, sin embargo, deben de aplicarse con mucho cuidado, pues en el intento de reducir la complejidad del árbol, se puede también reducir su precisión en los resultados.

El algoritmo CART utiliza un factor denominado: Gini, que indica la 'pureza' de cada uno de los nodos. Se puede decir que un nodo que posee a todos los elementos de una misma clase tiene una pureza del 100%, mientras tanto, un nodo que contiene elementos de 2 clases distintas y con la misma cantidad de elementos de una clase que de otra, tiene un nivel de pureza del 50%.

El cálculo del factor Gini se puede apreciar con claridad en la Figura 2.5, donde el valor p_i corresponde a la cantidad de elementos de clase 'i' en una determinada segmentación del árbol.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Figura 2.5 Ecuación para el cálculo del factor Gini (Will, 2016).

2.2.4.2 Bosques Aleatorios

Este algoritmo es más conocido como RFT (Random Forest Tree), es considerado como uno de los algoritmos de aprendizaje más certeros que existe. Permite resolver problemas de clasificación, pero a diferencia de otros algoritmos, utiliza una combinación de varios árboles de decisión con el uso de una técnica conocida como bagging (Hatcher, 2014).

El RFT funciona de la siguiente manera:

- Primero se computan diferentes árboles de decisión con la ayuda de subconjuntos aleatorios de datos (similar al algoritmo CART, se hace uso de la técnica RBS). Un ejemplo claro se puede ver en la Figura 2.6.
- Se evalúa la entrada con todos los diferentes árboles de decisión generados en el paso anterior.
- El resultado final es aquel que más árboles de decisión compartan.

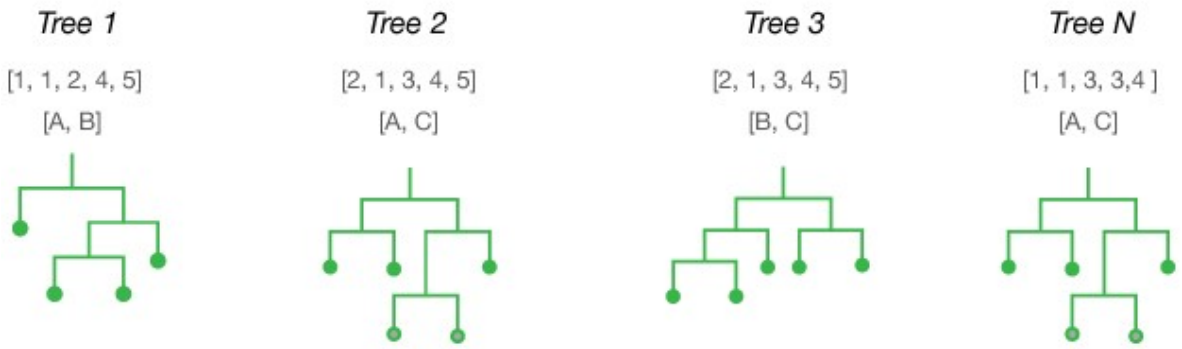


Figura 2.6 Diferentes árboles de decisión generados por el algoritmo RFT (Hatcher, 2014)

2.2.4.3 ID3

El algoritmo ID3, al igual que el algoritmo de CART, tiene como finalidad generar un árbol de decisión, sin embargo, la principal diferencia radica en que el ID3 lo logra basándose en

el principio de entropía, el cual, consiste en calcular la 'ganancia de información' al tomar una decisión o no, eligiendo siempre la opción que represente mayor ganancia (Sayad, 2012). En la Figura 2.7 se puede apreciar cómo funciona el ID3, mientras que la Figura 2.8 ilustra el principio de entropía utilizado por este algoritmo.

```
Id3(Ejemplos, Atributo-objetivo, Atributos )
  •Si todos los ejemplos son positivos devolver un nodo positivo
  •Si todos los ejemplos son negativos devolver un nodo negativo
  •Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo en Ejemplos

   En otro caso
        Sea A Atributo el MEJOR de atributos
        Para cada v valor del atributo hacer
              Sea Ejemplos(v) el subconjunto de ejemplos cuyo valor de atributo A es v
             •Si Ejemplos(v) esta vacío devolver un nodo con el voto mayoritario del Atributo objetivo de Ejemplos
              Sino Devolver Id3(Ejemplos(v), Atributo-objetivo, Atributos/{A})
```

Figura 2.7 Pseudocódigo correspondiente al algoritmo ID3 (Gómez Fernández, 2013)

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

Figura 2.8 Ejemplo de cálculo de entropía (Sayad, 2012).

Este algoritmo es considerado como no incremental, es decir, requiere de un conjunto fijo de instancias de entrenamiento. El ID3 solo maneja atributos discretos, además, las clases o grupos creados por ID3 son inductivos, es decir, a partir de un pequeño conjunto de instancias de entrenamiento, se espera que estas clases puedan clasificar también a nuevas instancias futuras.

2.2.4.4 C5.0

Un algoritmo que nace como una mejora al C4.5, así como de una extensión al ID3. El C5.0 utiliza el concepto de diferencia de entropía, de manera similar al ID3, esto le ayuda a decidir entre dividir un nodo o no (Rulequest, 2017).

El C5.0 trabaja de la siguiente manera:

- Primero que todo se verifican los casos base, llámense nodos hoja.
- Para cada atributo, se debe encontrar la diferencia de entropía o ganancia de información al dividirlo.
- Se opta por aquel atributo con el cuál se obtiene una ganancia mayor.
- Se procede a crear un nodo que divida el atributo seleccionado.
- Repetir los mismos pasos en las sub-listas de instancias generadas por la división del nodo.

La forma de trabajo del C5.0 es muy parecida a la del ID3, sin embargo, existen algunas diferencias, por ejemplo: el C5.0 utiliza técnicas de poda de árboles que permiten obtener modelos más reducidos e igual de eficientes, además, es capaz de manipular tanto atributos continuos como discretos mientras que el ID3 se limita únicamente a atributos discretos.

2.2.5 Sistemas Expertos

2.2.5.1 Redes Bayesianas

Las Redes Bayesianas son grafos acíclicos dirigidos (DAG) en donde los nodos son variables y los arcos representan relaciones de influencia causal entre los nodos. Los parámetros que se usan para representar la incertidumbre, no son más que las probabilidades condicionadas de cada nodo dado los diferentes valores que pueden tomar los nodos padres. En otras palabras, asumiendo un conjunto de variables {Xi, i = 1…, n} y que Pa (Xi) representa el conjunto de padres del nodo i, los parámetros de la red son las distribuciones condicionadas {P (Xi /Pa (Xi)), i=1..., n} (Castillo, Gutiérrez, & Hadi, 1997).

Para definir una Red Bayesiana, es necesario detallar:

- El conjunto de variables.
- El conjunto de enlaces entre variables, con los cuáles se debe formar una estructura DAG.
- Para cada variable, su probabilidad condicionada al conjunto de sus padres.

Las Redes Bayesianas utilizan el concepto de Independencia Condicional, principio que describe lo siguiente:

'Todo nodo es condicionalmente independiente de sus no descendientes, dado sus padres'

En probabilidad, dos acontecimientos R y B son condicionalmente independientes dado un tercer evento Y, si la ocurrencia o no ocurrencia de R junto con la de B se da en forma independiente dada Y. En otras palabras, R y B son condicionalmente independientes dado Y, si y sólo si el conocimiento que se tiene de Y provoca que el conocimiento sobre el estado de R no genere cambios sobre la probabilidad de que ocurra B, y de igual manera el conocimiento de si se produce B no proporciona información sobre la probabilidad de que ocurra R (Castillo, Gutiérrez, & Hadi, 1997).

Las Redes Bayesianas reciben su nombre a partir del modelo gráfico que se usa para representarlas (red) y del teorema pilar que se utiliza para generar este modelo: el teorema de Bayes. Este teorema, tal y como lo demuestra la Figura 2.9, utiliza la probabilidad condicional y las probabilidades a priori para poder calcular una probabilidad a posteriori dado cierto conjunto de observaciones.

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Donde:
$P(A_i)$ = Probabilidad a priori
$P(B/A_i)$ = Probabilidad condicional
$P(B)$ = Probabilidad Total
$P(A_i/B)$ = Probabilidad a posteriori

Figura 2.9 Teorema de Bayes (Juárez Ibujes, 2016).

A continuación, la Figura 2.10 muestra un pequeño ejemplo de una Red Bayesiana donde cada nodo cuenta con su tabla inicial de probabilidades, las cuáles eventualmente se verán alteradas ante la presencia de 'evidencia' u observaciones que permitan asignar valores concretos a uno o muchos de los nodos, y que permitirán retroalimentar y propagar las nuevas probabilidades a toda la red.

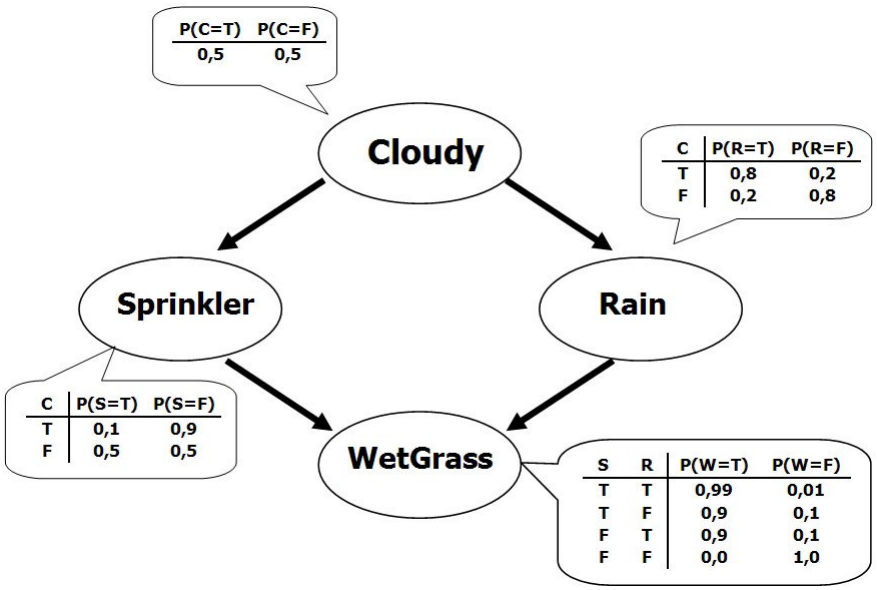


Figura 2.10 Ejemplo de probabilidades a priori y posteriori de una Red Bayesiana (Carvalho, 2015).

En la Tabla 2.2 se detallan algunas de las ventajas y desventajas que poseen las Redes Bayesianas respecto a otras técnicas de la Inteligencia Artificial.

Ventajas	Desventajas
Las relaciones son fácilmente entendibles por cualquier usuario. No se requiere un conjunto completo de entrada, puede manejar conjuntos incompletos. Permite dar una explicación sobre las causales de un resultado.	Necesita un enorme conjunto de datos probabilísticos. En ocasiones, puede resultar bastante complejo construir el grafo, porque no todas las dependencias pueden llegar a ser modeladas.

Tabla 2.2 Ventajas y Desventajas de las Redes Bayesianas (Elaboración Propia).

2.2.5.2 Reglas de Inferencia

Muchas veces nos encontramos con situaciones complejas gobernadas por reglas deterministas, por ejemplo: los sistemas de control de tráfico, sistemas de seguridad, transacciones bancarias (Sancho Caparrini, 2017).

Los sistemas basados en reglas son una herramienta eficiente para tratar este tipo de problemas. Las reglas deterministas constituyen la más sencilla de las metodologías utilizadas en sistemas expertos.

Se cuenta con una base de conocimiento, la cual contiene un conjunto de variables y reglas, así como se cuenta también con un motor de inferencia, que se encarga de obtener conclusiones con ayuda de las reglas y aplicando la lógica clásica.

Se debe entender por regla, una proposición lógica que relaciona dos o más objetos, e incluye dos partes: la premisa y la conclusión. Cada una de estas partes consiste en una expresión lógica con una o más afirmaciones objeto-valor conectadas mediante los operadores lógicos: y, o, no. Una regla se puede traducir normalmente como: 'Si premisa, entonces conclusión'.

El motor de inferencia es el componente encargado de gerenciar y controlar lógicamente el conocimiento almacenado en la base de conocimientos. El paradigma del motor de inferencia es la estrategia de búsqueda para producir el conocimiento demandado.

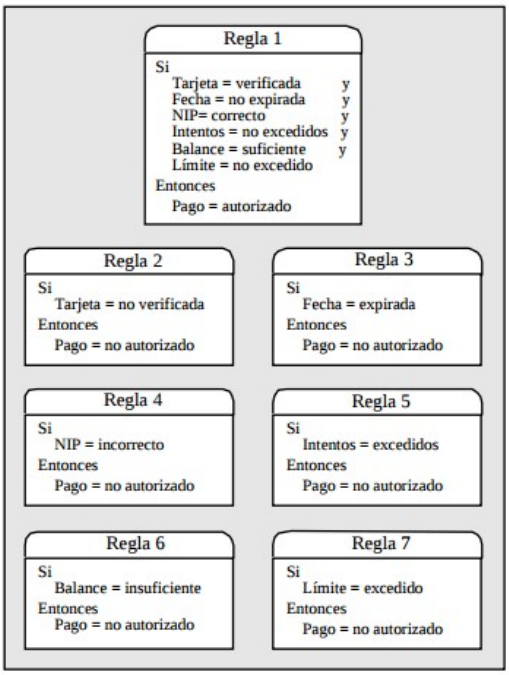


Figura 2.11 Ejemplo de reglas para sacar dinero de un cajero automático (Sancho Caparrini, 2017).

Existen 2 reglas de inferencia que suelen ser las más utilizadas por los sistemas expertos basados en reglas: el Modus Ponens y el Modus Tollens.

El Modus Ponens consiste en examinar la premisa de la regla, la cual, de ser cierta, permite que la conclusión pase a formar parte del conocimiento del sistema. Por otro lado, el Modus Tollens examina primero la conclusión, que, de ser falsa, se infiere que la premisa también lo es. La Figura 2.12 ilustra la forma de operar de cada uno de estos dos mecanismos de inferencia.

El rendimiento del motor de inferencia depende del conjunto de reglas en su base de conocimiento, dado que hay situaciones en las que es posible llegar a una conclusión utilizando un sub conjunto de las reglas, mientras no es posible utilizando otro.

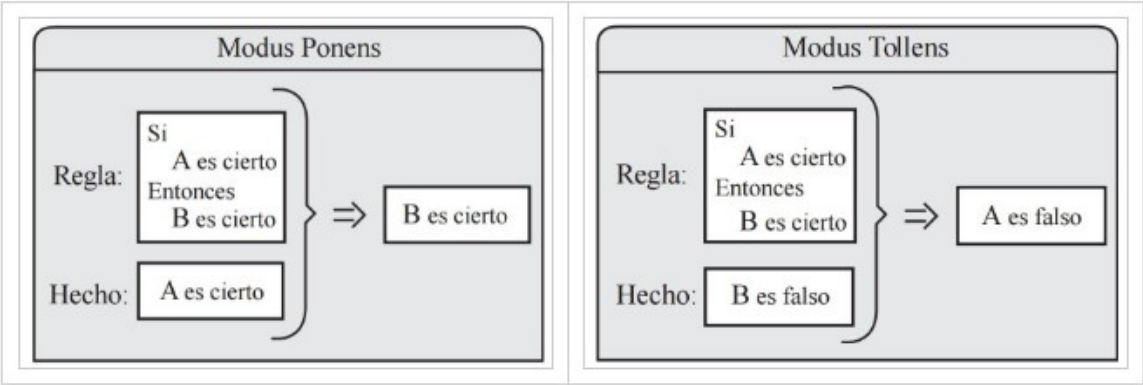


Figura 2.12 Modus Ponens y Modus Tollens (Sancho Caparrini, 2017).

A continuación, la Tabla 2.3 describe algunos de los pros y contras que poseen los Sistemas Expertos basados en reglas.

Ventajas	Desventajas
Cada regla representa una 'unidad de conocimiento', lo cual permite crear sistemas modulares. Todo el conocimiento es expresado en el mismo formato: 'reglas', esto facilita la construcción de la base de conocimientos.	Existe la posibilidad de caer en ciclos infinitos y contradicciones si no se definen las reglas apropiadamente. Hacer modificaciones de la base de conocimiento puede ser algo complicado.

Las reglas son un formato natural e intuitivo para expresar conocimiento de un dominio.	Algunos dominios son demasiado grandes o complejos, por lo que pueden llegar a ser necesarias millones de reglas.

Tabla 2.3 Ventajas y Desventajas de los Sistemas Expertos basados en reglas (Elaboración Propia)

2.2.6 Redes Neuronales

Las Redes Neuronales, más conocidas como ARN (por sus siglas en inglés), son modelos que intenta reproducir el comportamiento del cerebro basándose en el modelo biológico, es decir, en la estructura y funcionamiento del sistema nervioso (Introducción a las redes neuronales artificiales, s.f.).

Haciendo una analogía con el modelo biológico, las señales que llegan a la sinapsis se convierten en las entradas de la neurona artificial. Estas señales son procesadas con la ayuda de un parámetro denominado peso, asociado a la sinapsis respectiva. Dependiendo de la señal recibida, la neurona puede activarse como puede ser inhibida. Para determinarlo, se calcula el efecto total de todas las entradas, que finalmente es comparado con un valor llamado umbral. Si el efecto calculado es mayor al umbral, la neurona será activada, análogamente, si el efecto es menor, la neurona será inhibida. Una red neuronal está compuesta por múltiples neuronas como la que se muestra en la Figura 2.13. Cada una de las neuronas se organizan en capas, y existen 3 diferentes tipos de capas:

- **Capa de Entrada:** Encargada de recibir directamente los datos que provienen de fuentes externas a la red.
- **Capas Ocultas:** Son las capas internas de la red y no tienen contacto directo con el entorno exterior. Pueden ser interconectadas y distribuidas de diferentes maneras dependiendo del modelo de red que se elija. La forma cómo interactúan las capas ocultas tienen un alto impacto en el rendimiento y precisión de la red neuronal.
- **Capa de Salida:** Tiene contacto directo con el entorno exterior al igual que la capa de entrada, sin embargo, su única función es transferir la información producida por la red hacia el exterior.

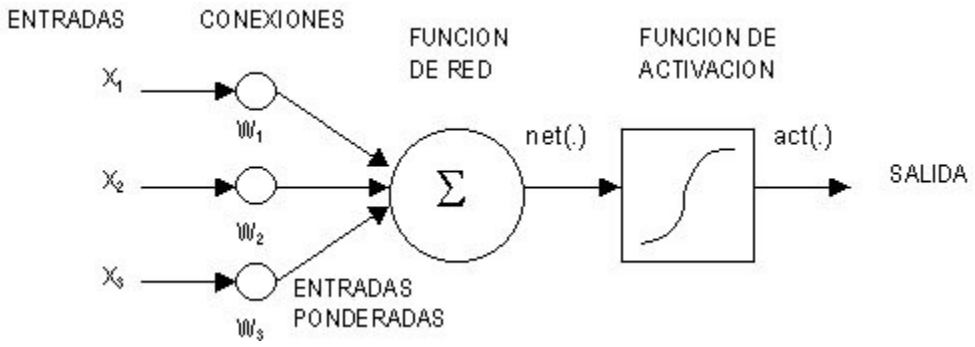


Figura 2.13 Estructura básica de una neurona (Introducción a las redes neuronales artificiales, s.f.)

La conectividad entre los nodos de una red está relacionada con la forma en que las salidas están canalizadas para servir de entrada hacia otras neuronas. Existen modelos de redes neuronales donde incluso la salida de una neurona puede convertirse en la entrada de sí misma creando un ciclo.

Las ARN son bastante apropiadas para aplicaciones en las que no se dispone, a priori, de un modelo identificable que pueda ser programado, pero se dispone de un conjunto básico de ejemplos de entrada. Asimismo, son altamente robustas tanto al ruido como a la disfunción de elementos concretos.

2.3 CASOS DE ÉXITO

2.3.1 Reconocimiento de Patrones en imágenes citológicas cervicales en 2D para la detección temprana del cáncer cervical (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009)

El análisis de imágenes es un proceso costoso que toma tiempo y es muy propenso a errores humanos, lo cual puede concluir en un diagnóstico o resultado erróneo. Esto fue lo que motivó a un grupo de investigadores de la Universidad Siwss German en Indonesia a desarrollar un sistema entrenado para reconocer patrones complejos y llevar a cabo un proceso de clasificación altamente preciso diferenciando entre células cancerígenas y no cancerígenas.

El trabajo de investigación se llevó a cabo en 5 pasos importantes:

- **Construcción de datos:** El experimento se realizó utilizando 235 ejemplos de células cervicales extraídas de una colección de exámenes de Papanicolaou realizados en el distrito de Tangerang, Indonesia durante el 2008. Todas las imágenes se obtuvieron en un formato JPEG. Durante la validación, todos los datos fueron diagnosticados por especialistas en citología del hospital Dharmais y validados en la clínica Kloster Paradiese, Alemania. El conjunto de entrenamiento estuvo comprendido por 100 datos, de los cuáles, 50 pertenecían a células normales y otras 50 a células cancerígenas.
- **Procesamiento de Imágenes:** Esta fase emplea cuatro pasos secuenciales, abarca conversión de imagen, mejora, segmentación y filtrado. Durante este proceso las imágenes 2D se convierten en formato de escala de grises. Trabajar utilizando el formato de escala de grises ofrece una gran ventaja, la cual consiste en que el sistema es capaz de reconocer los parámetros con facilidad debido a que se evita la confusión que puede causar la presencia de varios colores.
- **Extracción de Características:** En esta fase se extraen las características de la imagen, es decir, de la célula en estudio. Para la extracción de características de la imagen se procesa la imagen 2D aplicando la transformada de ondícula como se puede apreciar gráficamente en la Figura 2.14 y la Figura 2.15.

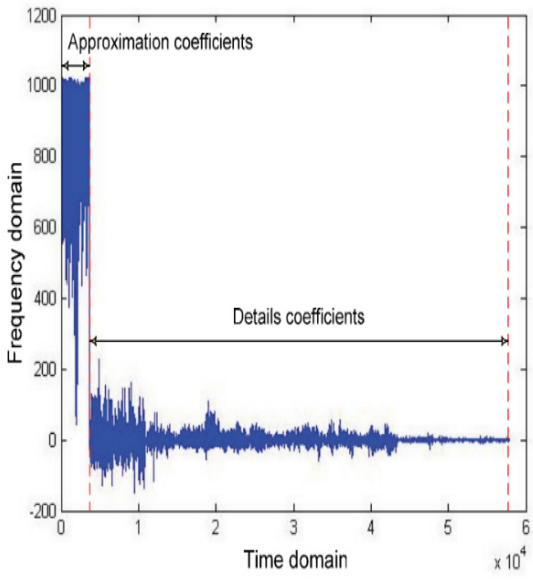


Figura 2.14. Función ondícula asociada a la extracción de características de una imagen. (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009).

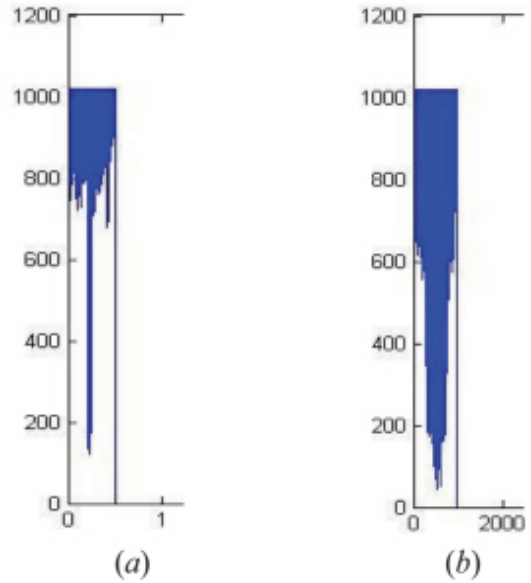


Figura 2.15. Aproximación de coeficientes de la función ondícula. (a) célula normal, (b) célula cancerígena (Suryatenggara, Ane, Pandjaitan, & Steinberg, 2009).

- **Reconocimiento de Patrones:** Basado en las observaciones previas, 4 patrones importantes fueron encontrados: el tamaño del núcleo y el citoplasma, cuya relación indica la madurez de la célula; la distancia entre célula puesto que las células cancerígenas tienden a andar juntas y formar colonias; la intensidad del color, ya que las células cancerígenas tienen un color ligeramente más oscuro que las células normales; la función ondícula, ya que las células cancerosas tienen un intervalo más amplio de valores de baja frecuencia dentro de sus coeficientes de aproximación que las células no cancerígenas.
- **Automatización y Clasificación:** Diferentes algoritmos como KNN, SVN y Naive Bayes fueron utilizados para desarrollar la etapa de automatización y clasificación. Los resultados obtenidos al aplicar las 3 técnicas mencionadas sobre los patrones y características pre procesadas fueron los siguientes: Generalmente una célula normal del cuello uterino tiene el área del núcleo más pequeña y el área del citoplasma más grande. En términos de la transformada de ondícula, estas características morfológicas de la célula están representadas por un patrón específico de los coeficientes de aproximación donde tiene un intervalo más estrecho de valores bajos

en el dominio de frecuencia. Mientras tanto, basándose en la intensidad del color, una célula normal tiene una mayor distribución dentro del intervalo de intensidad de color alto, que representa una proporción mayor de píxeles más brillantes en la imagen. Debido a la forma morfológica del núcleo, los coeficientes de aproximación de las ondas de una célula cancerosa tienen un intervalo más amplio de valores bajos en el dominio de la frecuencia, por lo tanto, los coeficientes son más distribuidos en los intervalos de clase de valores bajos. Además, en términos de intensidad de color, una célula cancerosa tiene principalmente mayor distribución en los niveles de baja intensidad, lo que significa una mayor proporción de píxeles oscuros en la imagen.

2.3.2 Aplicación de Redes Bayesianas Dinámicas para la predicción de cáncer cervical (Onísko, Druzdzel, & Austin, 2009)

El PCCSM es un modelo basado en redes bayesianas dinámicas que fue desarrollado y puesto a prueba con las herramientas SMILE (motor de inferencia) y GeNIe (entorno de desarrollo para modelos probabilísticos), ambos desarrollados por la universidad de Pittsburgh. Este modelo permite evaluar el nivel de riesgo de un tumor cancerígeno para de esta forma poder tomar decisiones correctas acerca del seguimiento y diagnóstico del paciente. El PCCSM está conformado por 19 variables en total, tal como se puede apreciar en la Figura 2.16, dentro de las cuales se cuenta con variables citológicas (relacionadas al resultado del Papanicolaou), histopatológicas (relacionadas a resultados de biopsias y cirugías) y el hrHPV DNA (variable referida al resultado del test del virus del papiloma humano) considerando también como relaciones temporales y dinámicas los arcos existentes entre las variables: Edad, Prueba HPV (prueba del papiloma humano) y el riesgo de tener cáncer de cervical.

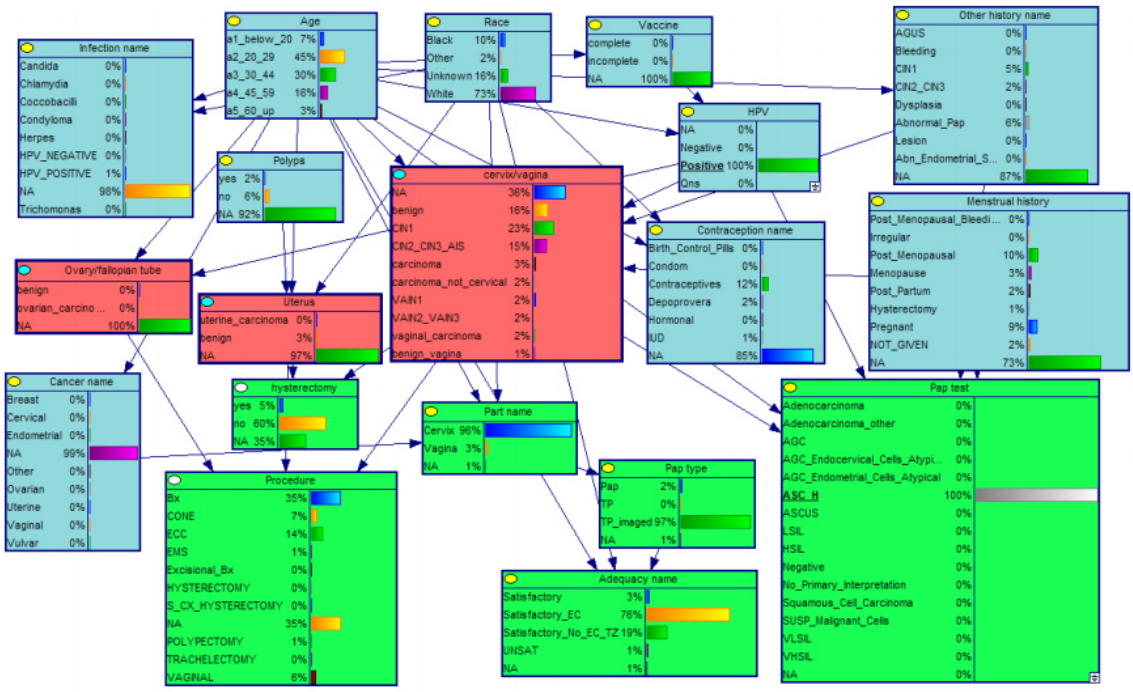


Figura 2.16. Representación gráfica del PCCSM con 19 variables (Onísko, Druzdzel, & Austin, 2009).

Haciendo uso de la Red Bayesiana Dinámica, se puede analizar en una gráfica cómo es que la tendencia o riesgo de cáncer de cuello cervical incrementa o disminuye con el paso del tiempo como consecuencia de una serie de eventos, lo cual ayuda a realizar estudios sobre la enfermedad en mención, y más aún, sobre cada caso clínico en particular logrando ver la evolución de la paciente. La Figura 2.17 muestra el gráfico generado por el PCCSM para el caso clínico de un paciente evaluado a lo largo de 15 años. La imagen permite apreciar claramente cómo durante el primer y el tercer año la paciente experimenta los momentos más riesgosos debido a un resultado anormal obtenido en un test de Papanicolaou y un resultado positivo en el test HPV; sin embargo, a partir del cuarto año se ve como el riesgo va disminuyendo gradualmente, esto debido a que se dio un retraso en el proceso de infección del virus del papiloma humano.

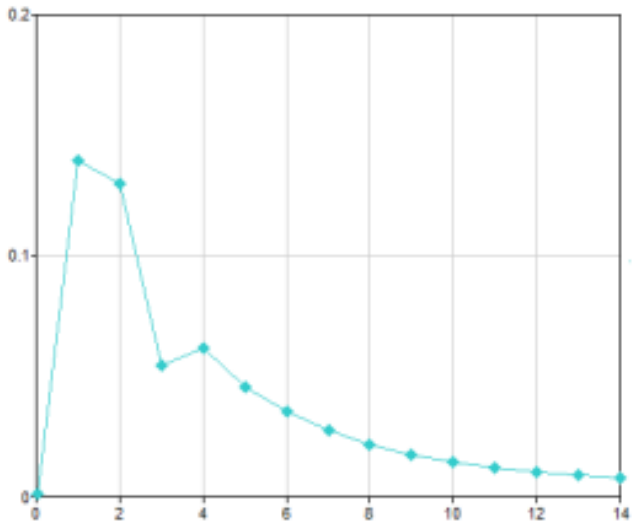


Figura 2.17. Resultados de riesgo de un paciente evaluado durante 15 años (Onísko, Druzdzel, & Austin, 2009).

Se utilizó en total 347 601 casos de pacientes previamente diagnosticadas con cáncer de cuello uterino para el proceso de entrenamiento del PCCSM y fueron 45 930 los utilizados para el proceso de validación que se enfocó principalmente en analizar cada uno de los casos individualmente verificando que correspondan a las opiniones de patólogos expertos.

En conclusión, el modelo PCCSM permite obtener resultados cuantitativos para computar estimaciones de riesgo utilizando diferentes variables que contribuyan al caso de estudio. Esta evaluación servirá como herramienta de control de calidad para la salud de los pacientes.

2.3.3 Sistema experto para diagnóstico temprano de cáncer de cuello uterino (Sanchez, 2012)

Los procesos de diagnóstico de cáncer de cuello cervical son muy ineficientes, existe retraso para la atención y entrega de resultados, retardo respecto al diagnóstico de los pacientes, carencia de personal de ginecología (en particular en la clínica Belén, Chiclayo, Perú) y otra serie de factores, como el hecho de que muchas mujeres no acuden al médico a realizarse un diagnóstico temprano de cáncer de cuello uterino, que hacen que esta enfermedad siga siendo una de las principales causas de muerte en mujeres. Es por ello que Lourdes Yosli planteó la implementación de un sistema experto que automatice el proceso de diagnóstico de cáncer de cuello uterino utilizando reglas de inferencia.

El motor de inferencia fue construido en base al árbol de dominio presentado en la Figura 2.18, el cual fue obtenido a través de entrevistas con un experto en el área de ginecología. El sistema fue pensado como una plataforma web para que esté abierto al uso de cualquier persona.

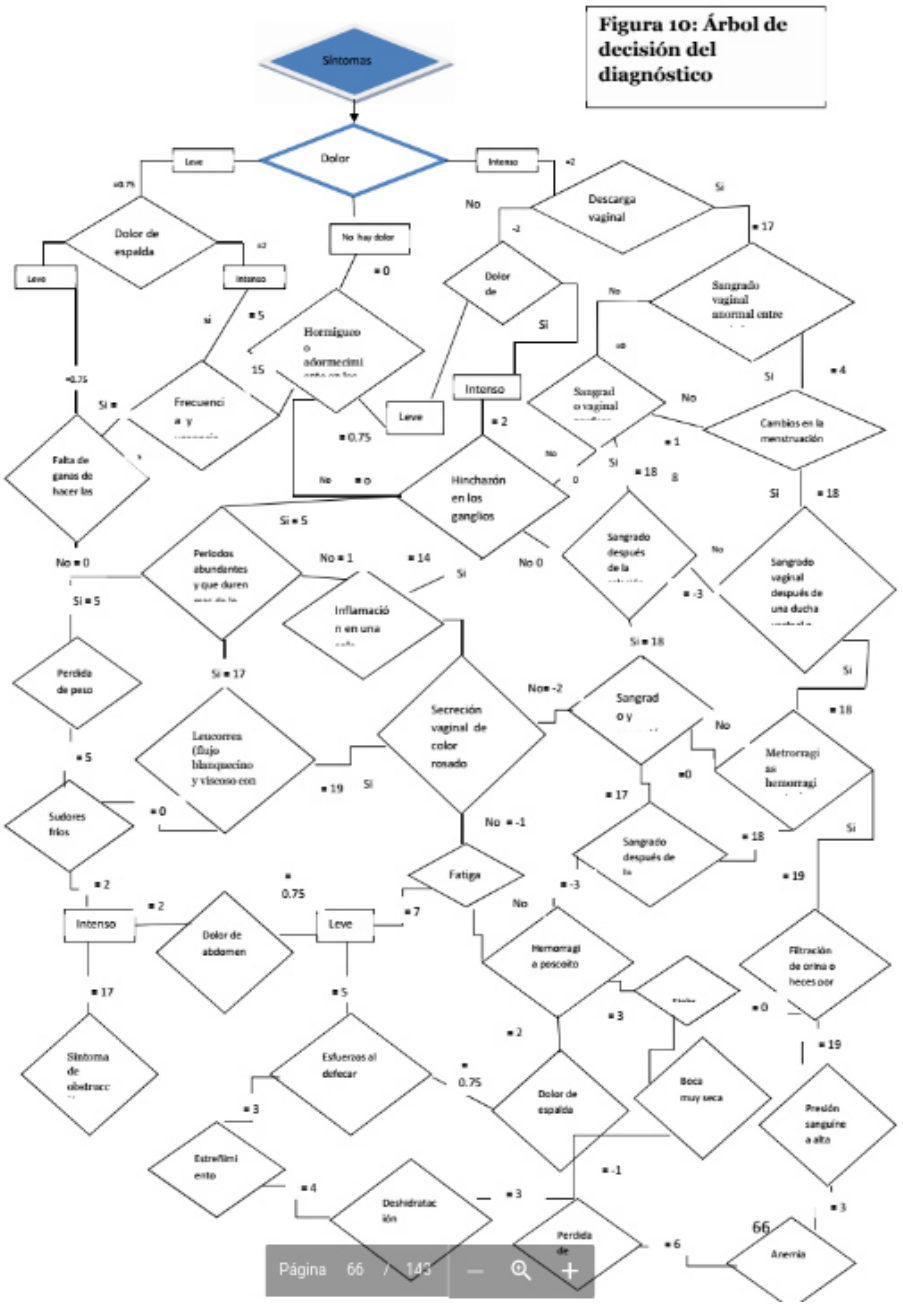


Figura 2.18. Árbol de dominio utilizado en las reglas de inferencia (Sanchez, 2012).

El software fue implementado satisfactoriamente en la Clínica Maternidad Belén. Este sistema experto basado en reglas de inferencia logró reducir el tiempo promedio que el ginecólogo se tomaba para dar un diagnóstico de cáncer de cuello uterino. El tiempo máximo que un doctor se tomaba eran de 45 minutos, mientras que, con el uso del sistema experto propuesto, los diagnósticos son realizados en aproximadamente 6 minutos. Además, la propuesta se puso a prueba con 113 casos clínicos pertenecientes a la misma clínica obteniendo una tasa de éxito increíblemente alta que asciende al 97%.

2.3.4 La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino (Pereira & Chamorro, 2012)

En este artículo se presentan los resultados del proyecto de investigación cuyo objetivo fue extraer patrones de supervivencia en mujeres con diagnóstico de cáncer invasivo de cuello uterino utilizando técnicas de minería de datos a partir de la información almacenada en el Registro Poblacional de Cáncer del Municipio de Pasto (Colombia), durante el periodo de 1998 a 2007. De acuerdo a los resultados obtenidos aplicando la técnica de clasificación basada en árboles de decisión, el tiempo de supervivencia de estas mujeres es mayor que 37 meses, contados a partir de la fecha de diagnóstico hasta la fecha última de observación de este estudio. Aplicando la tarea de Asociación se conocieron los principales factores socioeconómicos y clínicos asociados a la supervivencia de este grupo poblacional. El conocimiento generado permitirá soportar la toma de decisiones eficaces de los organismos gubernamentales y privados del sector salud en lo relacionado con el planteamiento de políticas públicas y programas de protección a las mujeres con esta enfermedad.

Se pueden distinguir 5 etapas importantes para llevar a cabo este proceso de investigación, las cuales fueron:

- **Etapa de selección de datos:** El objetivo de esta etapa fue obtener las fuentes internas y externas de datos que sirven de base para el proceso de minería de ellos.
- **Etapa de limpieza de datos:** Durante esta etapa se buscó obtener datos limpios, es decir, datos sin valores nulos o anómalos que permitan obtener patrones de calidad. Por medio de consultas SQL ad-hoc o a través de histogramas, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de

una tabla llamada CANCER, donde se encuentran los registros correspondientes a mujeres con cáncer de cuello uterino.

- **Etapa de transformación de datos:** La idea es transformar la fuente de datos en un conjunto listo para aplicar las diferentes técnicas de minería de datos. Para facilitar la extracción de patrones, se crearon en la tabla CANCER atributos adicionales a partir de los atributos que se tenían en un inicio.
- **Etapa de minería de datos:** Las tareas de minería de datos escogidas para el proceso de descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino fueron clasificación y asociación, teniendo en cuenta que a partir de los atributos identificados, se puede construir un modelo de clasificación que determine las características de las pacientes que viven y de las que mueren y además se pueden identificar relaciones no explícitas entre los atributos del conjunto de datos pertenecientes a las mujeres que sobreviven por medio de asociaciones.
- **Etapa de interpretación y evaluación de resultados:** En esta etapa se interpretan los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto.

La Figura 2.19 muestra el resultado de la clasificación utilizando la herramienta Weka, mientras que en la Figura 2.20 se muestra el resultado de haber efectuado la asociación con la misma herramienta.

Los patrones más representativos de supervivencia en mujeres con cáncer invasivo de cuello uterino que fueron descubiertos son:

- Si el número de meses de vida transcurridos a partir de la fecha de diagnóstico hasta el 2007 es mayor a 37, entonces la mujer se consideró sobreviviente. El 42% de los 507 casos de mujeres con cáncer invasivo de cuello uterino se clasifican de esta manera, mientras el 63% de las mujeres vivas cumplen con este patrón.
- Las mujeres que no cumplieron el período de sobrevivencia tuvieron el siguiente patrón: no ser madre cabeza de familia, no estar clasificada en el SISBEN ni pertenecer a un régimen de salud. El 32.5% de los 507 casos de mujeres con cáncer

invasivo se clasificó de esta manera, mientras el 95.4% de las mujeres muertas cumple el mismo patrón.

```
weka.classifiers.trees.J48 -C 0.7 -M 20
=== Classifier model (full training set) ===
J48 pruned tree
------------------
nmeses <= 37
|   cabezaflia <= 0
|   |   nivelsisben = 1: VIVO (50.0/20.0)
|   |   nivelsisben = 7: MUERTO (165.0/38.0)
|   |   nivelsisben = 2: VIVO (33.0/6.0)
|   |   nivelsisben = 3: VIVO (4.0/1.0)
|   |   nivelsisben = 4: VIVO (1.0)
|   cabezafamilia > 0: VIVO (42.0/4.0)
nmeses > 37: VIVO (212.0/15.0)

Number of Leaves  :         7
Size of the tree  :         10
```

Figura 2.19. Resultados de clasificación en Weka (Pereira & Chamorro, 2012).

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 0 -C
0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c 19
Relation:     QueryResult-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.NumericToNominal-R6-
weka.filters.unsupervised.attribute.Discretize-B5-M-1.0-
R5-weka.filters.unsupervised.attribute.Remove-R4
Instances:    334
Attributes:   14
              region
              comuna
              edaddx
              nmeses_2007
              cabezafamilia
              regimen
              nivelsisben
              fuenteagua
              tipovivienda
              estrato
              escolaridad
              estadocivil
              ocupacion
              discapacidad
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.2 (67 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 16
```

Figura 2.20. Resultados de asociación en Weka (Pereira & Chamorro, 2012).

2.3.5 Predicción de recurrencia en pacientes con cáncer de cuello uterino utilizando MARS y Clasificación (Chang, Cheng, Lu, & Liao, 2013)

El cáncer cervical es, hoy en día, una enfermedad que cuenta con diversos métodos para ser detectada a tiempo y recibir tratamiento con buenos resultados, sin embargo, algo que aún no se ha logrado es tratar eficientemente el problema de la recurrencia de esta enfermedad en pacientes ya diagnosticados, lo que dio lugar a la presente investigación.

Para llevar a cabo el estudio se utilizó un conjunto de datos proporcionado por el Registro de Tumores del Hospital Universitario Médico de Chung Shan para verificar la viabilidad y efectividad de C5.0 y MARS. Cada paciente en el conjunto de datos contiene 12 variables predictoras: edad, tipo de célula, grado del tumor, tamaño del tumor, pT (primary tumor), pStage (estado del tumor), margen quirúrgico, metástasis ganglionar (LNM), número de fracciones de radioterapia, resumen objetivo de radioterapia, espacio linfo-vascular (LVSI), secuencia de terapia locorregional y terapia sistémica.

El conjunto de datos consta de casos correspondientes a 168 pacientes. Entre ellos se seleccionaron aleatoriamente 118 conjuntos de datos con respecto a la proporción de pacientes recurrentes y no recurrentes como la muestra de entrenamiento (estimando los parámetros de los correspondientes modelos de clasificación construidos) mientras que los restantes 50 serán retenidos como la muestra de prueba (evaluando la capacidad de clasificación de los modelos construidos).

En el modelo de clasificación C5.0, primero deben seleccionarse las variables predictoras (o independientes). Dos variables independientes significativas se incluyeron en el modelo final C5.0, las cuáles fueron pT y el resumen objetivo de radioterapia. Los resultados obtenidos fueron del 96,0% (tasa de éxito).

Para modelar el modelo de clasificación MARS, todas las doce variables predictoras se utilizan como entradas. Aplicando este modelo, la tasa promedio de clasificación correcta obtenida fue del 86,00%.

El modelo C5.0 tiene la mejor capacidad de clasificación en términos de la tasa de clasificación promedio correcta, supera al modelo de MARS y por lo tanto proporciona una alternativa eficiente en la realización de las tareas de clasificación de cáncer cervical. Con el

fin de evaluar la robustez y el rendimiento de los métodos C5.0 y MARS, se probó utilizando 10 etapas de pruebas independientes. Con base en los hallazgos, el modelo C5.0 no sólo genera el mejor resultado de clasificación, sino que también puede utilizarse para seleccionar variables independientes importantes para la clasificación del cáncer de cuello uterino, las cuales pueden proporcionar información útil para el tratamiento.

Después de 10 etapas de pruebas, las variables seleccionadas fueron: pStage, pT, tipo de célula y resumen objetivo de radioterapia, dando a entender que son éstos los factores más influyentes para obtener un buen resultado de clasificación en el presente caso de estudio. La Figura 2.21 detalla más el resultado luego de las 10 etapas de pruebas.

Model Runs	{1-1}		{2-2}		Overall	
	C5.0	MARS	C5.0	MARS	C5.0	MARS
1	100.00	88.24	87.50	81.25	96.00	86.00
2	91.67	97.22	100.00	85.71	94.00	94.00
3	95.00	90.00	80.00	70.00	96.00	86.00
4	89.47	89.47	100.00	75.00	92.00	86.00
5	91.89	97.30	92.31	76.92	92.00	92.00
6	94.87	87.18	81.82	90.91	92.00	88.00
7	84.38	96.88	94.44	33.33	88.00	74.00
8	97.14	94.29	80.00	53.33	92.00	82.00
9	95.12	95.12	100.00	77.78	96.00	92.00
10	88.89	88.89	92.86	78.57	90.00	86.00
Average	92.05	92.46	91.27	72.28	92.44	86.60

Figura 2.21. Resultados obtenidos en cada una de las 10 etapas de pruebas independientes (Chang, Cheng, Lu, & Liao, 2013).

Como conclusión el estudio demuestra que las técnicas de árboles de decisión tienen una alta tasa de éxito al solucionar problemas de predicción, para este caso en particular, la recurrencia del cáncer cervical.

2.3.6 Identificación de regiones anormales en la zona cervical utilizando imágenes del examen de colposcopía (Liang, Zheng, Huang, Milledge, & Tokuta, 2013)

El cáncer cervical representa una de las principales causas de muerte para las mujeres en países en desarrollo y la predicción se perfila como una de las mejores formas de atacar el problema, motivo por el cual nace este estudio donde se propone un algoritmo capaz de identificar las regiones anormales de la zona cervical a partir de una secuencia de imágenes extraídas del examen de colposcopía.

Durante la colposcopía se aplica un ácido llamado ácido acético que provocará ciertos cambios de color en la zona cervical luego de ser aplicado. Este método es considerado uno de los mejores indicadores para detectar regiones cancerígenas y pre-cancerígenas, sin embargo, la zona cervical es bastante limitada y a veces resulta complicado detectar y rastrear ciertas características de la imagen obtenida de forma consistente.

El algoritmo propuesto dividirá la zona cervical en 3 regiones importantes:

- El epitelio escamoso (se conserva de color rosado luego de aplicado el ácido)
- El epitelio columnar (región oscura entre el endometrio y el epitelio escamoso)
- Región blanca (región que se torna blanca luego de aplicado el ácido)

Además. la reflexión especular es grande y se mantiene cambiante debido a los movimientos producidos durante la colposcopía, por lo que se incluyó un módulo encargado de remover la reflexión especular. La Figura 2.22 detalla el algoritmo propuesto la implementación de este módulo.

Para remover la reflexión especular se utilizó la técnica de interpolación bilineal, de esta forma las secuencias de imágenes extraídas pueden ser estudiadas con el menor ruido posible. Se utilizó la técnica SVM para realizar el proceso de predicción con las secuencias de imágenes luego de realizada la segmentación de regiones sobre la zona cervical la cual se puede apreciar en la Figura 2.23.

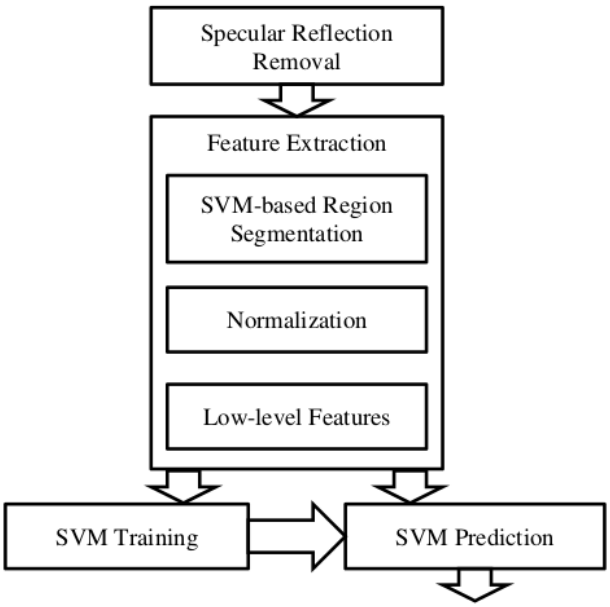


Figura 2.22. Algoritmo propuesto para remover la reflexión especular (Liang, Zheng, Huang, Milledge, & Tokuta, 2013).

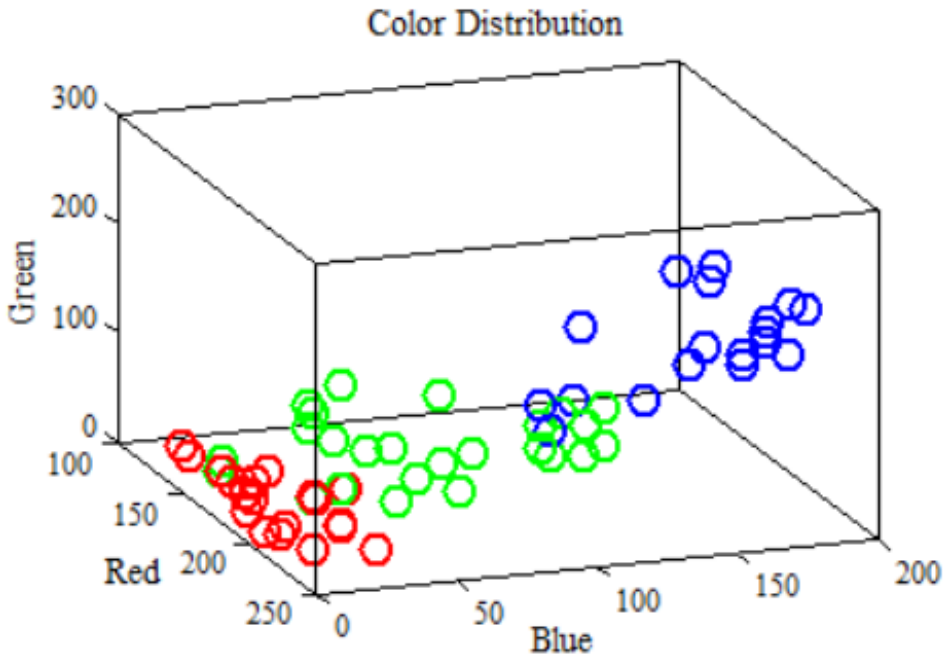


Figura 2.23. Segmentación de regiones Rojo (regiones con sangre), Verde (regiones oscuras), Azul (otras regiones) (Liang, Zheng, Huang, Milledge, & Tokuta, 2013).

La prueba de validación se realizó sobre 48 conjuntos de imágenes pertenecientes a 12 pacientes con diferentes diagnósticos respecto al cáncer cervical. Cada conjunto conteniendo diferentes imágenes de la zona cervical de diferentes ángulos, zoom y brillo. Finalmente, los resultados dieron a conocer que el 94.6% fue la mayor tasa de éxito obtenida utilizando la SVM con una función kernel lineal, la cual, a pesar de la más simple, demostró mejor comportamiento para este caso de estudio que las funciones RBF y Polinomial.

En conclusión, el algoritmo logra satisfactoriamente predecir cuándo un paciente puede desarrollar o tiene cáncer cervical, además de no requerir imágenes 100% precisas.

2.3.7 Prevención y detección de cáncer utilizando técnicas de Minería de Datos (Ramachandran, Girija, & Bhuvaneswari, 2014)

Ramachandran, Girija y Bhuvaneswari desarrollaron y pusieron a prueba un modelo para detección temprana y prevención de cáncer que desarrollaron utilizando minería de datos y técnicas de clasificación, clustering y predicción con el fin de identificar posibles pacientes con cáncer y detectar la predisposición de una persona de ser diagnosticada con cáncer antes de tener que pasar por exámenes clínicos y de laboratorio que son costosos y a la vez demandan tiempo.

Se hizo uso de diversas técnicas como Árboles de decisión y Clustering utilizando el algoritmo K-means para la implementación del algoritmo propuesto que se puede apreciar en la Figura 2.24.

El árbol de decisión es generado utilizando una técnica de división binaria recursiva y permite extraer patrones frecuentes de todo el conjunto de datos que tengan relación significativa con los conjuntos de datos de cáncer y no cáncer de tal forma que los patrones estén bien distinguidos para separarse entre estos dos conjuntos. El árbol de decisión asignará un peso a cada atributo y determinará cuáles son los patrones frecuentes. La Figura 2.25 muestra el árbol de decisión de salida.

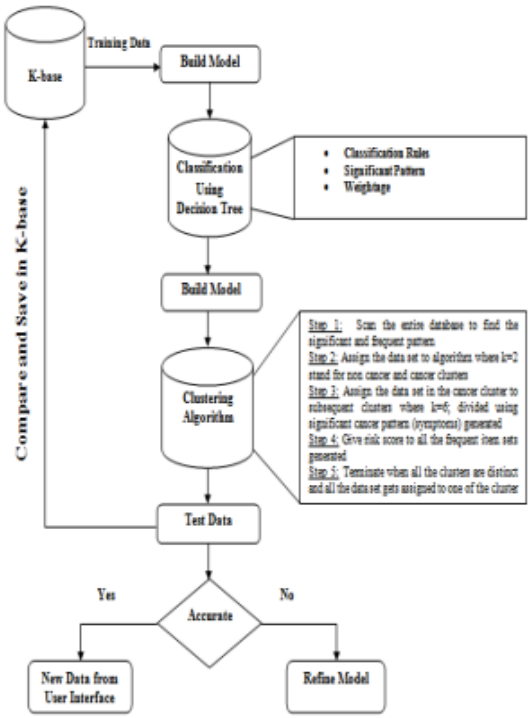


Figura 2.24. Algoritmo propuesto: Árboles de decisión, Clustering y K-Means (Ramachandran, Girija, & Bhuvaneswari, 2014).

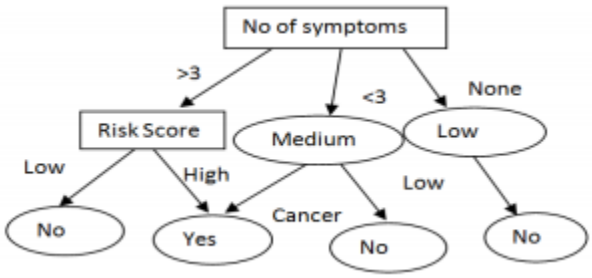


Figura 2.25. Árbol de decisión generado por división binaria recursiva (Ramachandran, Girija, & Bhuvaneswari, 2014).

El siguiente paso viene dado por el Clustering, para el cual se hizo uso del algoritmo K-means, ahora los datos se agrupan en un número de clases donde cada clase es identificada por una única característica basada en los patrones extraídos por el árbol de decisión. Los pesos son introducidos como entrada al algoritmo K-means para agruparlos y dividirlos en grupos de cáncer y no cáncer. Dentro de estos dos clusters, se crearán 6 sub-clusters donde cada uno de estos representará un tipo de cáncer (pulmón, cervical, mama, estómago, oral, leucemia) al cuál el objeto de datos será asignado dependiendo de los síntomas presentes.

De todos los datos que se pudo recolectar, se utilizó el 75% para construir el modelo de clasificación y agrupación mientras que los datos restantes se utilizaron para efectuar pruebas. Los datos recolectados pertenecen a personas que presentan y no presentan cáncer y además constan de más de 30 atributos tales como edad, estado civil, síntomas relacionados con el cáncer, peligros, antecedentes familiares de cáncer, etcétera.

El 99% del total de datos utilizado para la validación (746) fueron clasificados correctamente, por lo que se concluye que el sistema de predicción es una solución real, rápida, efectiva y de bajo costo para la detección temprana del cáncer.

2.3.8 Redes Bayesianas para el apoyo en el diagnóstico de pacientes evaluados con el test de Papanicolaou (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014)

Un conjunto de investigadores de la Universidad Nacional Tecnológica de Atenas, Grecia, desarrolló un sistema basado en redes bayesianas para apoyar en el diagnóstico de pacientes que se sometieron a las pruebas del Papanicolaou y obtuvieron resultados ambiguos. Para la implementación del sistema se utilizó el software Weka (plataforma para el aprendizaje automático y minería de datos escrito en Java y distribuido como software libre bajo la licencia GNU-GPL).

Para llevar a cabo la implementación, se construyó una Red Bayesiana haciendo uso del algoritmo de aprendizaje K2 y las siguientes variables: Resultado del examen de Papanicolaou de acuerdo al sistema Bethesda, resultados del examen HPV DNA referente al virus del papiloma humano, resultados del examen NASBA para identificar otros tipos de HPV, resultados del Flujo de Isometría para identificar expresiones de alto riesgo de HPV y la expresión inmunocitoquímica de p16 para detectar mutaciones que podrían aumentar el riesgo de desarrollar cáncer. La Figura 2.26 muestra varias relaciones importantes entre las variables utilizadas como por ejemplo la relación entre el HPV y NASBA donde se señala que primero debe existir una infección (detectada por HPV) y luego la integración del virus (detectada por el NASBA), además, la progresión a CIN2+ (detectada por la histología) no es factible sin la integración del virus detectado por NASBA. La Figura 2.27 muestra algunas evidencias con sus respectivas probabilidades.

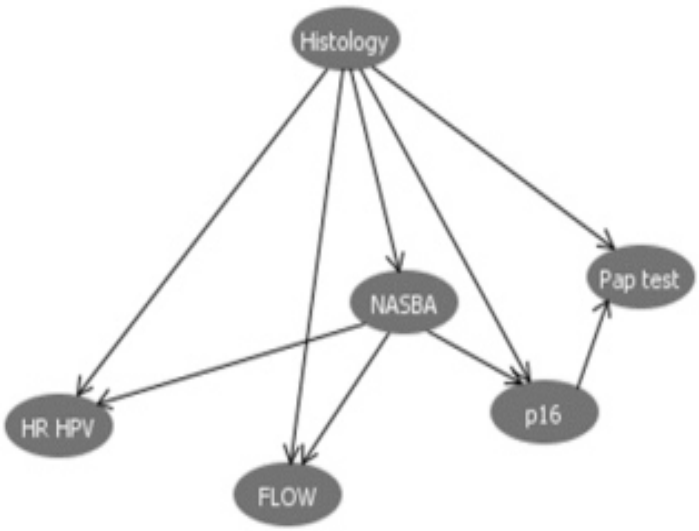


Figura 2.26. Representación gráfica de la Red Bayesiana utilizada por Bountris et al (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014).

Scenario	Pap Test	HR HPV DNA	NASBA	Flow Cytometry	p16	BN Results (Posterior Probabilities %)
1	ASCUS	Positive	Unknown	Positive	Positive	Negative=4, CIN1=22, CIN2/3=66, CxCa=8
2	ASCUS	Negative	Unknown	Unknown	Unknown	Negative=63, CIN1=34, CIN2/3=3, CxCa=1
3	ASCUS	Positive	Positive	Negative	Negative	Negative=14, CIN1=64, CIN2/3=16, CxCa=1
4	LSIL	Negative	Unknown	Positive	Unknown	Negative=7, CIN1=73, CIN2/3=17, CxCa=3
5	LSIL	Positive	Negative	Positive	Negative	Negative=1, CIN1=70, CIN2/3=28, CxCa=1
6	LSIL	Positive	Positive	Positive	Positive	Negative=2, CIN1=41, CIN2/3=51, CxCa=6
7	HSIL	Negative	Negative	Negative	Unknown	Negative=31, CIN1=49, CIN2/3=16, CxCa=4

Figura 2.27. Probabilidades asociadas a algunas evidencias (Bountris, Tsirmpas, Koutsouris, & Haritou, 2014).

En total, un conjunto de 740 datos fue analizado, usándose el 70% (512 casos) de los datos para el entrenamiento de la red y el 30% (228 casos) restante fue utilizado para el proceso de validación de la red bayesiana.

La aplicación del modelo propuesto dio resultados prometedores lo que sugiere que es posible mejorar la precisión del diagnóstico y apoyar el triaje de ASCUS/LSIL. De acuerdo a los resultados, la red bayesiana demostró los resultados más balanceados en términos de sensibilidad (84.3%), especificidad (94.9%), PPV (82.7%) y NPV (95.5%).

2.3.9 Detección y clasificación de cáncer cervical utilizando análisis de texturas (Soumya, Sneha, & Arunvinodh, 2016)

Existen diferentes herramientas para el diagnóstico de cáncer, tales como los rayos X, la tomografía y la resonancia magnética que con la ayuda de técnicas de procesamiento de imágenes pueden detectar la enfermedad.

En el caso de estudio que desarrollaron Soumya, Sneha y Arunvinodh, proponen utilizar una técnica de clasificación aplicada al resultado de una prueba de resonancia magnética para poder identificar la etapa de cáncer cervical en la que se encuentra el paciente (Ver Figura 2.28)

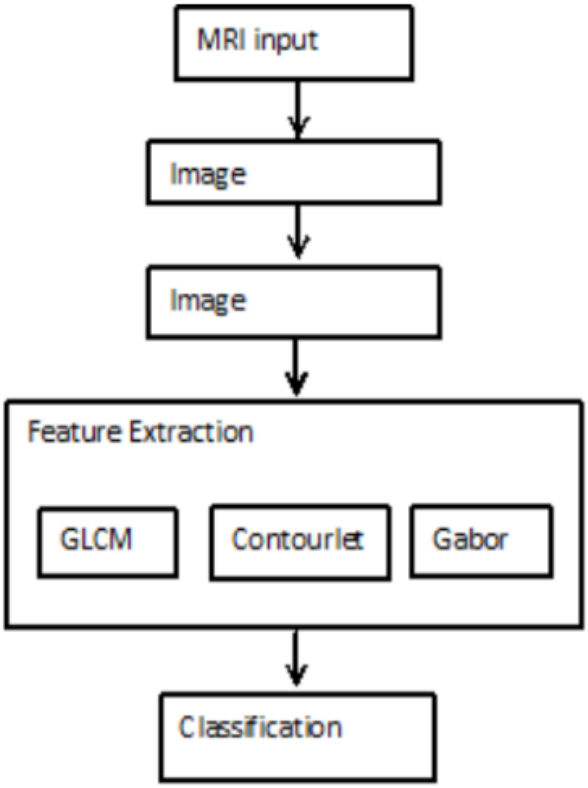


Figura 2.28. Esquema del sistema propuesto por Soumya et al (Soumya, Sneha, & Arunvinodh, 2016).

Las dos etapas principales fueron:

- **Procesamiento de Imágenes:** Durante esta etapa, se utilizaron las imágenes obtenidas por resonancia magnética para pre-procesarlas y obtener características útiles para el sistema de predicción. Se empezó realizando una limpieza de la imagen

utilizando la corrección Gamma seguido de la aplicación de la técnica de Otsu para segmentar la imagen y reconocer las regiones de interés. Finalmente se procede con la extracción de características de las imágenes, proceso que se realizó con la aplicación de 3 métodos distintos: Matriz de Concurrencia (GLCM), transformación de Contourlet y filtro de Gabor.

- **Clasificación:** Se construyeron modelos de clasificación SVM no lineales basados en características de textura de segundo orden y características de transformación de los tumores. Las características obtenidas por la transformación de Contourlet y las características de Gabor, se utilizan para la predicción de la salida. Sin embargo, las características estadísticas de segundo orden basadas en el contraste, la correlación, la energía y la homogeneidad se utilizan significativamente para predecir el resultado de pre-tratamiento de imágenes de resonancia magnética (MRI) de tumores de cáncer cervical. Los modelos basados en transformadas con mejor desempeño tuvieron una efectividad del 81% para las MRI axiales ponderadas en T1, 82% para T2 axial y 83% para imágenes sagitales ponderadas T2, algo mejor que los modelos basados únicamente en factores clínicos. Por lo tanto, las características de textura superaron las características de transformación.

Una pequeña comparación entre el nuevo modelo y el antiguo se puede observar en la Figura 2.29.

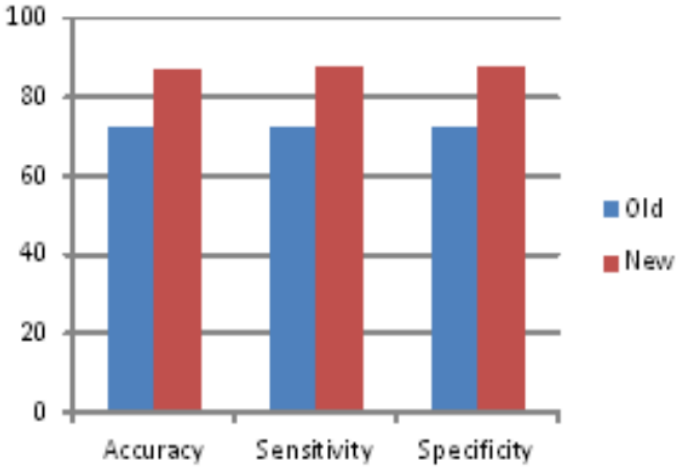


Figura 2.29. Comparación del nuevo modelo propuesto utilizando procesamiento de imágenes versus el modelo actual utilizando solo factores clínicos (Soumya, Sneha, & Arunvinodh, 2016).

2.3.10 Predicción del cáncer de cuello uterino mediante inducción híbrida (Vidya & Nasira, 2016)

Vidya y Nasira proponen en su trabajo de investigación, utilizar un conjunto de técnicas utilizadas dentro de la minería de datos para predecir el estado (benigno o maligno) de las células del tejido cervical. Para lograr el objetivo del caso de estudio, se estudió el comportamiento de diferentes algoritmos como el algoritmo de CART y el RFT. Además, se contó con diversos conjuntos de datos proveído por el NCBI (National Center for Biotechnology Information) donde cada uno de ellos contenía información de 61 variables numéricas extraídas de un análisis genético previamente realizado.

Luego de haber realizado las pruebas con ambos algoritmos respectivamente se implementó una técnica híbrida aplicando una combinación entre el RFT y la técnica de K-Means. En todas las pruebas, las salidas fueron presentadas en formas de árboles de decisión y se utilizaron 100 registros de los cuales 60 pertenecen al conjunto de entrenamiento y 40 pertenece al conjunto de pruebas.

- **Aplicación del algoritmo de CART**: Para la aplicación del algoritmo de CART se utilizó el factor GINI que busca las mejores características en cada nodo interno para luego crear una decisión. Este método alcanzó un 83.7% de tasa de éxito en su predicción.
- **Aplicación de RFT**: El algoritmo RFT es introducido con dos finalidades: la primera, para construir una regla de predicción y segundo, para evaluar y clasificar las variables adquiriendo la capacidad de predecir un resultado. El resultado en términos de tasa de éxito fue del 93.54% y la forma como se aplicó el RFT fue la siguiente:
  - ✓ Se tomaron muestras vectoriales aleatorias de los registros obtenidos de la NCBI (muestras vectoriales de 61 variables).
  - ✓ El algoritmo genera un bosque de árboles por cada una de las muestras vectoriales aleatorias seleccionadas.
  - ✓ Se utilizan los puntos de estimación obtenidos del bosque de árboles para realizar la predicción sobre cáncer maligno y benigno.
- **Aplicación de RFT + K-Means**: Los predictores RFT podrían conducir a una medida de disimilitud entre las observaciones sobre datos no etiquetados. La idea principal

es construir un predictor RFT para distinguir los datos clínicos observados de los conjuntos de datos generados, pre procesador por un algoritmo de clustering como K-Means debido a que RFT puede manejar variables homogéneas y heterogéneas con mucha facilidad que otras variables. Este método híbrido propuesto alcanzó un 97.77% de tasa de éxito.

A continuación, la Figura 2.30 muestra a detalle cada uno de los resultados.

No.	Methods	Prediction in percentage
1	CART	83.87%
2	RFT	93.54%
3	RFT with K-mean	96.77%

Figura 2.30. Resultados obtenidos luego de la aplicación de CART, RFT y RFT + K-Means (Vidya & Nasira, 2016).

2.3.11 Método de identificación de cáncer cervical sobre imágenes de histología basado en las características de textura y el área de lesión (Wei, Gan, & Ji, 2017)

Wei, Gan y Ji plantean el problema de un enfoque automatizado para detectar el cáncer cervical para el cual proponen mejorar la precisión del reconocimiento de la enfermedad. Las imágenes de histología del área cervical son necesarias y juegan un papel fundamental en el reconocimiento del cáncer y en la propuesta que consiste de 4 pasos importantes:

- Pre-procesamiento de las imágenes de histología para reducir el impacto provocado por el ruido presente, así como el impacto en la posterior extracción de características precisas.
- Las imágenes son particionadas y agrupadas en 10 imágenes verticales y se utiliza la Matriz de Co-ocurrencia de Nivel de Gris (GLCM, por sus siglas en inglés) para poder obtener información sobre la textura del área en análisis (contraste, correlación, entropía, uniformidad, energía y más).
- La imagen es segmentada utilizando el algoritmo de clustering K-means. Cada una de las 10 imágenes verticales son divididas en 3 capas.
- Finalmente, se procesan las características obtenidas por GLCM y las características obtenidas en cada una de las capas que contienen áreas lesionadas haciendo uso de una SVM.

A continuación, la Figura 2.31 muestra gráficamente la propuesta de los autores en su trabajo de investigación.

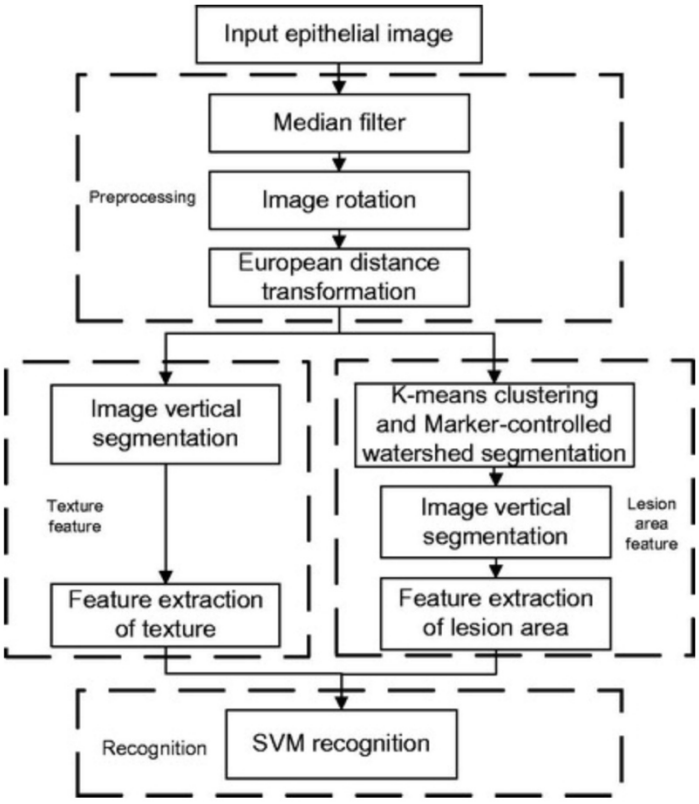


Figura 2.31. Propuesta desarrollada por Wei, Gan y Ji (Wei, Gan, & Ji, 2017).

A pesar de utilizar un conjunto bastante pequeño de datos, los resultados fueron bastante positivos. Según los autores, en un trabajo previo utilizando únicamente características obtenidas por GLCM el modelo alcanzó el 60% de precisión, de la misma forma, utilizando únicamente las características obtenidas de las áreas con lesión se alcanzó un 75%, mientras que combinando ambos conjuntos de características, se alcanza una precisión del 90%, un resultado que si bien es muy alentador, los autores toman con bastante discreción debido a la pequeña cantidad de datos utilizada para llegar a la conclusión.

2.3.12 Aplicación de Deep Learning para la clasificación de imágenes obtenidas por colposcopía (Sato, y otros, 2018)

En el 2018, en Japón, Sato y otros autores realizaron un trabajo de investigación que consistía en demostrar la eficiencia de Deep Learning aplicado a la práctica clínica ginecológica. Un

total de 485 imágenes divididas de la siguiente manera: 142 imágenes para displasia grave, 257 para CIS (carcinoma cervical in situ) y 86 imágenes para IC (invasive carcinoma).

Uno de los más grandes retos que tuvieron que afrontar fue el overfitting, que suele ser un problema bastante común en este tipo de aplicaciones, la gran discrepancia que existía entre la curva de entrenamiento y la curva de validación sugirió que se había producido un overfitting, probablemente causado por el pequeño número de imágenes incluidas, ya que normalmente se suelen preparar entre 500 y 1000 imágenes para cada clase. Por este motivo fue necesario recurrir a métodos que ayuden a evitar el overfitting causado por la limitación en el número de imágenes incluidas, como por ejemplo la regularización L2 y L1.

La arquitectura de la Red Neuronal utilizada para este trabajo se describe en la Figura 2.32.

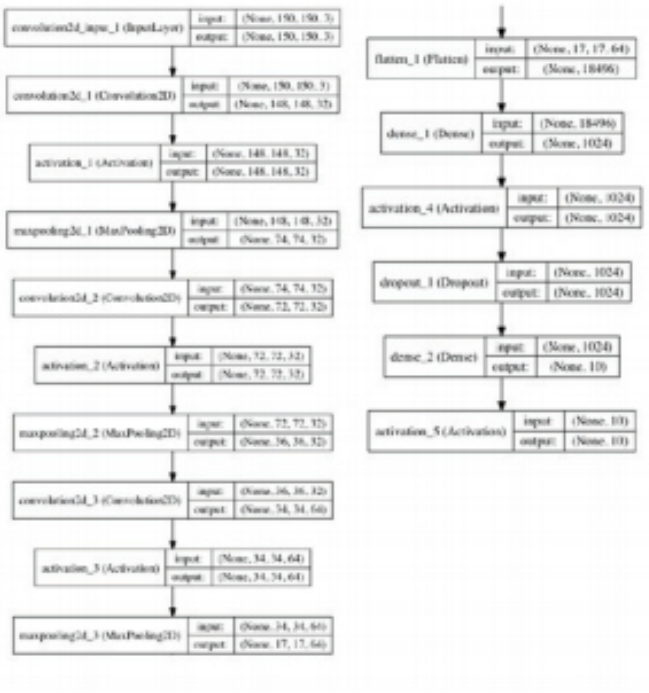


Figura 2.32. Arquitectura de la Red Neuronal utilizada para evaluar las imágenes de la zona cervical (Sato, y otros, 2018).

La precisión alcanzada con el conjunto de datos y el modelo propuesto alcanzó un 50%. Si bien es cierto este resultado fue considerado preliminar debido a la poca cantidad de datos, sugiere que los médicos e investigadores no necesitan ser unos expertos en inteligencia artificial o Machine Learning para utilizar Deep Learning debido a la gran cantidad de herramientas que existen hoy en día para procesar datos y generar conclusiones, como la

utilizada en la investigación: TensorFlow. Los resultados sugieren también, que mientras más información relevante se pueda obtener en las imágenes, incluyendo aquellas extraídas de exámenes como la colposcopía, deben ser siempre almacenadas digitalmente para su uso a futuro, ya que podrían enriquecer el desarrollo de sistema con Inteligencia Artificial como el que fue planteado por los autores.

3 CAPÍTULO III: MÉTODO PROPUESTO – REDES BAYESIANAS

En el presente capítulo se realiza una comparación entre todos los métodos analizados previamente en el **Capítulo II. Estado del Arte** dando paso a la justificación del método propuesto y descripción de la metodología a seguir.

3.1 JUSTIFICACIÓN

Antes de proceder a realizar la comparación, es necesario definir una serie de criterios y características que deben tener cada una de las técnicas revisadas previamente y que son relevantes para el caso de estudio. Los criterios a tener en consideración son:

- **Naturaleza de Datos:** Indica los tipos de datos que pueden ser usados como entrada por la técnica evaluada, estos pueden ser: continuos, discretos o ambos.
- **Cantidad de datos de Entrenamiento:** Hace referencia a la cantidad de datos de entrenamiento que requiere la técnica en análisis para adquirir el conocimiento o información necesaria para desempeñarse correctamente.
- **Interpretabilidad**: Representa con qué facilidad se pueden interpretar las causales del resultado obtenido por la técnica aplicada.
- **Velocidad:** Nos ayuda a medir la rapidez con la que el método en estudio procesa los datos de entrada para obtener el resultado.
- **Precisión:** Permite conocer qué tan exacto es el resultado del método aplicado en comparación con los resultados reales o resultados esperados.
- **Manejo de Ruido:** Expresa la capacidad del método para manejar datos ruidosos, es decir, datos corruptos que necesiten un pre-procesamiento para poder ser entendidos y analizados correctamente.
- **Área de Dominio:** Representa qué tan limitada es la técnica respecto a la complejidad de los diferentes dominios de conocimiento en los cuáles puede ser aplicada.
- **Nivel de complejidad:** Mide el esfuerzo humano necesario para modelar, implementar y validar la técnica en mención.

A continuación, en la Tabla 3.1 se muestran los valores que cada una de los criterios puede tomar, teniendo asociado a cada uno de estos un valor numérico que representa el puntaje de

cada valor. Al final, estos puntajes serán determinantes para efectuar la comparación entre todas las técnicas en evaluación.

CRITERIO	VALOR	DESCRIPCIÓN	PUNTAJE
Naturaleza de datos	Continuo Discreto	El método sólo admite que el dato de entrada sea discreto o que sea variable.	1
	Ambos	El método es adaptable a ambos tipos de datos.	2
Cantidad de datos de entrenamiento	Bajo	No se requiere de muchos datos de entrenamiento para alcanzar un alto rendimiento.	2
	Alto	Se requiere de un alto número de datos de entrenamiento para alcanzar un alto rendimiento.	1
Interpretabilidad	No Interpretable	El método no cuenta con un razonamiento simbólico y representación semántica.	0
	Compleja	El método es descriptivo y requiere cierta interpretación.	1
	Fácil	Se presentan los resultados de manera visual o al menos de manera que su interpretación sea muy clara.	2
Velocidad	Baja	El método requiere de un alto costo computacional, incluso si la cantidad de datos a manipular no es alta, por lo que los tiempos de respuesta son lentos.	1
	Alta	El método no siempre requiere de un alto costo computacional, dependiendo de la cantidad de datos a manipular, los tiempos de respuesta pueden ser incluso inmediatos.	2
Precisión	Baja	El método tiene una precisión <= 70%.	1
	Media	La precisión está entre 70% y 85%.	2
	Alta	El método tiene un indicador de precisión mayor a 85%	3
Manejo de ruido	Bajo	El método tiende a cometer grandes imprecisiones cuando recibe datos con alta presencia de ruido.	1
	Medio	Significa que la técnica es capaz de reconocer algunos datos ruidosos, sin embargo, no es constante y su tasa de precisión puede verse afectada notablemente.	2
	Alto	Se maneja eficientemente cualquier dato recibido, almacenado o editado sin que esto afecte en gran medida la precisión del resultado.	3
Área de dominio	Baja	La construcción no requiere ningún dominio de conocimiento.	3

	Media	Se requiere una delimitación en el dominio de trabajo para la construcción.	2
	Alta	Se necesita conocimiento completo del dominio que abarca el sistema.	1
Nivel de complejidad	Baja	Los algoritmos utilizados son sencillos e intuitivos, con una curva de aprendizaje alta, y no requieren experiencia previa para ser implementados.	3
	Media	La implementación de los algoritmos tienen una curva media de aprendizaje, no son complicados de implementar, pero si requieren experiencia previa para su buen entendimiento.	2
	Alta	El método utiliza algoritmos complejos que requieren de una alta curva de aprendizaje y son difíciles de implementar si no se cuenta con experiencia previa.	1

Tabla 3.1 Valores y puntajes de los criterios de comparación (Elaboración propia).

La Tabla 3.2 demuestra que las técnicas con mayor puntaje total obtenido en base a los criterios definidos anteriormente son: las Redes Bayesianas y el algoritmo de árboles de decisión RFT. Sin embargo, los árboles de decisión + RFT alcanzan un puntaje bajísimo en el atributo Número de Datos, mientras que las Redes Bayesianas obtienen la mejor puntuación para ese criterio en particular. Además, las Redes Bayesianas no necesitan de un conjunto de entrada completo, sino que es posible utilizar como entrada incluso una sola variable de todo el conjunto de variables utilizadas en la red.

Criterio / Técnica	SVM	CART	RFT	C5.0	ID3	Redes Bayesianas	Reglas de Inferencia	Redes Neuronales	Algoritmos Genéticos	K-Means	KNN
Naturaleza de Datos	2	2	2	2	2	2	1	1	1	2	2
Cantidad de datos de entrenamiento	1	1	1	1	1	2	2	1	2	1	1
Interpretabilidad	1	2	2	2	2	2	2	0	1	1	1
Velocidad	1	1	1	1	1	2	2	2	2	2	1
Precisión	3	2	3	3	2	2	2	3	2	3	3
Manejo de Ruido	3	2	3	2	2	3	1	3	2	1	2
Área de Dominio	3	3	3	3	3	2	1	3	1	2	2
Nivel de Complejidad	1	2	2	2	1	2	3	2	1	3	3
TOTAL	15	15	17	16	14	17	14	15	12	15	15

Tabla 3.2 Benchmarking de las diferentes técnicas analizadas (Elaboración Propia)

Existe evidencia de investigaciones donde se utilizaron algunas de las técnicas mencionadas y fueron comparadas con los resultados de utilizar Redes Bayesianas, un claro ejemplo es el estudio realizado en el 2008 por la Universidad Politécnica de Madrid (Correa Valencia, Bielza, Pamies Teixeira, & R. Alique, 2008), donde los autores señalan que las Redes Bayesianas obtienen mejores resultados que un calificador basado en Redes Neuronales el cual mostró una precisión del 94.8% frente al 96.3% alcanzado por las Redes Bayesianas. Se comprobó además que esta técnica necesita un tiempo increíblemente menor al necesario para construir un modelo basado en Redes Neuronales que requiere de 12.69 segundos mientras que las Redes Bayesianas necesitan tan solo 0.08 segundos (resultados comprobados haciendo uso de una computadora Dell Dimension a 3Gz y 1.5GB).

De manera similar, en el 2016, la Universidad de Castilla realizó una investigación para comparar los resultados de aplicar Redes Bayesianas y SVM. La Figura 3.1 muestra la precisión obtenida por cada uno de los métodos utilizados, las cuatro primeras filas corresponden a variaciones de las SVM puestas a prueba con 5 escenarios diferentes, mientras que las filas siguientes corresponden a variaciones de las Redes Bayesianas. El último modelo correspondiente a una Red Bayesianas construida a partir del algoritmo K2 es el cual obtuvo mayor precisión en todos los escenarios (Rubio, Martínez-Gómez, Flores, & Puerta, 2016).

Accuracy for all classifiers when facing scenarios A–E. Highest accuracy by scenario is in bold.

Classification model	Scenario A	Scenario B	Scenario C	Scenario D	Scenario E
SVM_{lin}	0.499	0.522	0.465	0.662	0.449
SVM_{pol}	0.464	0.433	0.404	0.617	0.353
SVM_{rad}	0.464	0.433	0.404	0.617	0.353
SVM_{sig}	0.464	0.433	0.404	0.617	0.353
NB_c	0.649	0.625	0.575	0.592	0.497
NB_d	0.652	0.617	0.546	0.626	0.529
TAN	0.748	0.707	0.695	0.752	0.726
BN_{K2}	**0.816**	**0.761**	**0.823**	**0.885**	**0.844**

Figura 3.1. Resultados obtenidos por las SVM y las Redes Bayesianas (Rubio, Martínez-Gómez, Flores, & Puerta, 2016).

Por otro lado, en Irán, la Universidad de Ciencias Médicas publicó una investigación en la cual se reunió evidencia suficiente (de diferentes estudios realizados entre el año 2005 y

2015) para demostrar que las Redes Bayesianas son una técnica altamente confiable y tiene un rendimiento bastante alto. La Figura 3.2. Evidencia recolectada entre el 2005 y 20015 sobre la aplicación de Redes Bayesiana (Langarizadeh & Moghbeli, 2016). La Figura 3.2 muestra a detalle la evidencia recolectada por los autores.

Por todas las razones anteriormente mencionadas en el Benchmarking, y los casos de éxito analizados, la técnica elegida para implementar este caso de estudio fue las Redes Bayesianas.

Reference & Year	Subject Study	Illness	Number Of Variables	Performance measure of NBN (p-value0.01)	comparison
(2014) [5]	35,605 patients with lung cancer	Brain metastasis from lung cancer	6 Variables	accuracy:82.83% Sensitivity:80.84% specificity:84.59%	BN:accuracy:82% Sensitivity:83.28% specificity:80%
(2008) [11]	Data of 142 brain tumor patients	brain tumor	96 attributes	Accuracy: 84% Specificity:87% Sensitivity:80%	BN:accuracy:80% Sensitivity:73% specificity:85%
(2011) [20]	1700 patients	Prostate Cancer	4 variables	AUC = 66.2%	TAN: AUC = 58%
(2011) [9]	3866 patient	Minor head trauma	17 attributes	Sensitivity:95% Specificity:95% AUC: 95%	PIPPER: Sensitivity:75.5% Specificity:76.9% AUC: 84.1%
(2013) [22]	2318 patients	glaucoma severity	6 variables	Accuracy values greater than 80%	----------
(2007) [17]	210 high-risk women	Breast Cancer	6 features	the ROC curve (AUC) : 67.5%	----------
(2013) [23]	119 Chinese patients with (DKD) & 554 without DKD	diabetic kidney disease	10 clinical attributes	Accuracy:84%	Partial least squares regression (PLS): Accuracy: 71%
(2010) [24]	830 patients	dialysis in ill patients	2 input variables	AUC: 85.5%	SVM: AUC: 83.3%
(2008) [16]	128 patients	BC	13 variables	AUC :88.4% accuracy: 81.8% sensitivity: 75% specificity & PPV: between 83% and 86%	Logistic Regression (LR) :AUC :79.4% accuracy: 76.1% sensitivity: 75% specificity: 83%
(2007) [25]	169 patients	acute appendicitis	9 Variables	ROC analysis showed: BN model provided the most reliable & accurate results.	NBN works better than LR & artificial neural network (ANN)
(2013) [26]	240 patients	AE.	42 Variables	Accuracy: 68%,	Decision Tree (DT): Accuracy: 64.1%
(2011) [10]	1411 patients	AD.	312 318 (single nucleotide polymorphisms (SNPs)) SNP	AUC: 59%	LR: AUC:61.3%
(2012) [27]	40 patients	Toothache	14 pain parameters (P=14).	Accuracy :72%	----------
(2010) [28]	population of 558 Italian SSc.	SSc.	19 Variable	accuracy:76.9% sensitivity:72.2% specificity:81.6%	LR: accuracy:75.5% sensitivity:69.4% specificity:81.6%
(2006) [29]	1086 patients	anesthesia	11 Preoperative and intraoperative characteristics	AUC: 57% accuracy: 77%; sensitivity :18.3%, specificity:95.7%, PPV: 57.6% NPV: 78.6%.	LR: ROC curve: 66.9% accuracy: 64.2%; sensitivity :62.5%, specificity:64.7%, PPV: 36.1% NPV: 84.4%.
(2013) [30]	583 patients	LV	12 Attributes	accuracy: 82.16%; sensitivity :82.35%, specificity:83%,	NN: accuracy: 79%; sensitivity : 77.54%, specificity: 75.83%,
(2007) [31]	59 patients	AE.	5 Variables	Sensitivity: 85% Specificity: 78% Accuracy: 82% AUC: 88%	DT: Sensitivity: 79% Specificity: 94%
(2012) [32]	987 patients	CAD	113 Variables	AUC: 78%	SVM: AUC: 75%
(2013) [33]	26 adult asthma patients	AE.	20 Variables	Sensitivity: 80% Specificity: 77% Accuracy: 77%	SVM: Sensitivity: 84% Specificity: 80% Accuracy: 80%
(2013) [34]	45 subjects	dementia in Parkinson's disease	4 Variables	Sensitivity: 92.33% Specificity: 100% Accuracy: 93.33%	Filter Selection NB: (FSNB) Sensitivity: 86% Specificity: 100% Accuracy: 93.33%
(2015) [35]	345 type 2 diabetic patients	type 2 diabetic	7 features	Sensitivity: 41.03% Specificity: 84.96% Accuracy: 62.61%	SVM: Sensitivity: 50% Specificity: 78.72% Accuracy: 60.87%
(2007) [36]	2,949 subjects	CAD	11 variables	Sensitivity: 88.3% Specificity: 88.6% Accuracy: 88.4%	NN (MLP) Sensitivity: 87.1% Specificity: 85.3% Accuracy: 86.2%
(2012) [37]	227 healthy newborn infants	neonatal jaundice	----------	AUC: 88%	Bayes Net: AUC: 87%

Table 2. Summary of important factors in NBN predictors.PPV: positive predictive value; NPV: negative predictive value

Figura 3.2. Evidencia recolectada entre el 2005 y 20015 sobre la aplicación de Redes Bayesianas (Langarizadeh & Moghbeli, 2016).

3.2 METODOLOGÍA

Tal y como se puede apreciar en la Figura 3.3, la metodología consiste de tres pasos principales; el primero es la selección del conjunto de datos, seguido del tratamiento de datos que consiste en asegurarnos que el conjunto de datos esté listo para ser utilizado y finalmente la construcción del modelo probabilístico.

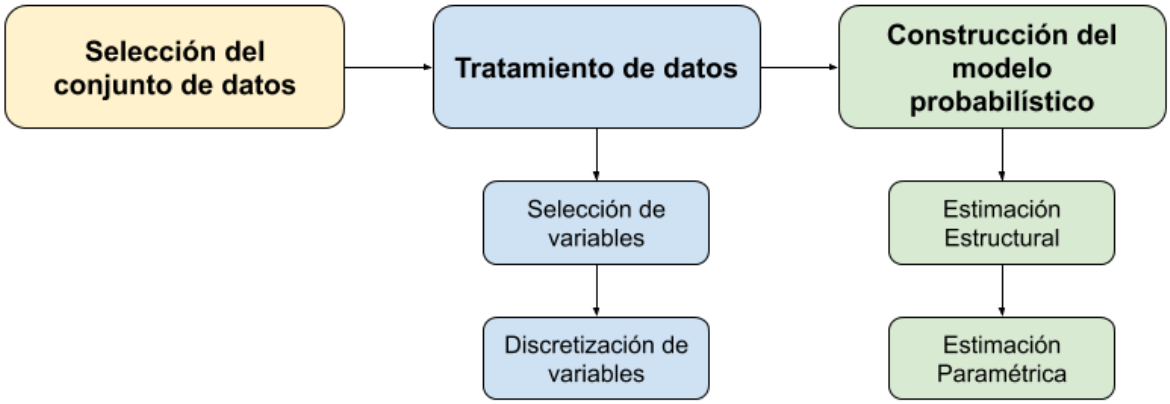


Figura 3.3. Representación gráfica de la metodología (Elaboración Propia)

3.2.1 FASE 1: SELECCIÓN DEL CONJUNTO DE DATOS

Uno de los problemas que gran parte de las investigaciones tienen que afrontar, es el de seleccionar un conjunto de datos. La verosimilitud, confiabilidad y coherencia de un conjunto de datos es vital para poder alcanzar buenos resultados en la investigación. Es por eso que en esta fase el objetivo principal es identificar una fuente lo suficientemente confiable a ser utilizada como fuente de datos, principalmente, histórico de pacientes que hayan sido diagnosticadas previamente con cáncer de cuello uterino tanto positivo como negativo. El experto en el área de dominio será el encargado de determinar si el conjunto de datos es confiable o no, basándose en su experiencia y alto conocimiento en el campo sobre el cual se está realizando el estudio, en este caso en particular, Ginecología.

3.2.2 FASE 2: TRATAMIENTO DE DATOS

Una vez se tiene un conjunto de datos confiable, es necesario trabajar en él para definir las variables que serán soportadas por el modelo probabilístico. Esta fase está divida a su vez en dos pequeñas sub fases:

3.2.2.1 SELECCIÓN DE VARIABLES

Esta sub fase consiste en analizar el conjunto de datos para identificar cuáles son las variables relevantes para el caso de estudio. Un conjunto de datos con información principalmente clínica de muchos pacientes puede contener atributos que no tienen relevancia con el caso de estudio, así como puede contener atributos que sí lo son, pero solo se conocen para una muy pequeña porción de todas las instancias, lo cual nos puede llevar a descartar la variable.

Nuevamente, el apoyo del experto es de suma importancia en esta etapa ya que, gracias a su conocimiento sobre el área de dominio, se podrá fácilmente conocer qué variables son importantes y de cuáles podemos prescindir.

Como resultado de esta sub fase se tendrá una lista completa de las variables a ser utilizadas en la implementación del modelo probabilístico.

3.2.2.2 DISCRETIZACIÓN DE VARIABLES

Luego de haber identificado cuáles son las variables a utilizar es importante clasificarlas en 2 tipos: variables discretas y variables continuas. Una variable discreta siempre es numérica, pero dentro de un intervalo siempre tiene un número finito de posibles valores, sin embargo, una variable continua es todo lo contrario, ya que puede tomar una cantidad infinita de valores dentro de un rango. Como parte de esta sub fase se busca definir una serie de intervalos para cada variable, en el caso de las variables discretas sólo si es estrictamente necesario para agrupar valores utilizando un criterio en común, y de manera obligatoria para las variables continuas ya que, al haber infinita cantidad de posibles valores, es necesario agruparlos para poder cuantificarlos correctamente.

Como resultado de esta sub fase se tendrá una lista completa de las variables, ahora discretas, cada una con sus posibles valores.

3.2.3 FASE 3: CONSTRUCCIÓN DEL MODELO PROBABILÍSTICO

Esta fase consiste en construir el modelo probabilístico haciendo uso del método propuesto: las Redes Bayesianas. Esta fase se divide en 2 pequeñas sub fases, las cuales son particular del método propuesto y se detallan a continuación.

3.2.3.1 ESTIMACIÓN ESTRUCTURAL

La estimación estructural consiste en, básicamente, definir la estructura topológica del grafo acíclico dirigido que dará vida a la Red Bayesiana. El objetivo principal de esta sub fase es obtener las relaciones entre cada una de las variables, en otras palabras, poder identificar qué variables influyen sobre cuáles otras.

El resultado de esta sub fase será un grafo, donde cada nodo represente una variable y las aristas representen la relación causal entre ellas.

3.2.3.2 ESTIMACIÓN PARAMÉTRICA

Una vez se tiene definida la estructura topológica de la Red Bayesiana, se necesita calcular las tablas de probabilidad de cada una de las variables. La Estimación Paramétrica es la sub fase que asignar estos valores a cada nodo, su objetivo principal es establecer los valores de probabilidad iniciales en cada nodo de forma que el modelo esté listo para propagar información cuando se requiera.

El resultado de esta sub fase será el mismo grafo de la sub fase anterior, pero donde cada nodo tendrá asociada una tabla de probabilidad que describe su comportamiento cuando se tiene evidencia de alguno de los nodos que influyen sobre él.

3.3 EJEMPLO UTILIZANDO EL MÉTODO PROPUESTO

A continuación, se procederá a mostrar un ejemplo (Puga, 2012) con cada una de las fases descritas anteriormente, que consiste en construir una Red Bayesiana capaz de diagnosticar gripe dado los síntomas de un paciente, asumiendo que solo existen dos tipos de gripe: Gripe común y Gripe A.

Selección de del conjunto de datos y Tratamiento de datos

Para poder llevar a cabo la implementación utilizando una Red Bayesiana y dar solución a este problema, se utilizará una base de datos de ejemplo con solamente 20 registros, tal y como se puede observar en la Figura 3.4.

Caso	Dolor_de_Cabeza	Problemas_Respiratorios	Enfermedad
1	SI	NO	Gripe_Comun
2	SI	NO	Gripe_Comun
3	NO	SI	Gripe_A
4	SI	NO	Gripe_Comun
5	SI	NO	Gripe_Comun
6	SI	SI	Gripe_A
7	SI	NO	Gripe_Comun
8	SI	NO	Gripe_Comun
9	SI	NO	Gripe_Comun
10	SI	NO	Gripe_Comun
11	NO	NO	Gripe_Comun
12	SI	SI	Gripe_A
13	SI	NO	Gripe_Comun
14	SI	NO	Gripe_Comun
15	SI	SI	Gripe_A
16	SI	NO	Gripe_Comun
17	SI	NO	Gripe_Comun
18	SI	NO	Gripe_Comun
19	SI	NO	Gripe_Comun
20	SI	NO	Gripe_Comun

Figura 3.4. Conjunto de datos a utilizar en el ejemplo (Puga, 2012).

Fácilmente se puede notar que, para este caso, no existen variables de tipo continuo y las variables discretas puede tomar a lo mucho dos valores, motivo por el cual no es necesario realizar la discretización.

Construcción del Modelo

Para comenzar con la Estimación Estructural se utilizará el proceso manual, es decir, se definirá la estructura de la RB en base al conocimiento de un grupo de expertos. Para el caso del ejemplo, se puede generar un modelo bastante simple que consiste en una RB divergente, también conocido como modelo de causa común, tal y como se muestra en la Figura 3.5.

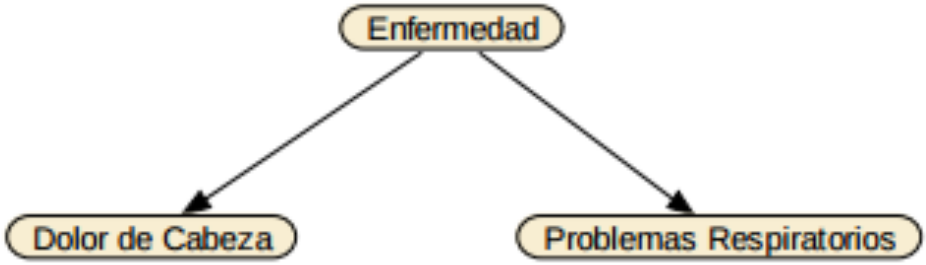


Figura 3.5. Estructura de la Red Bayesiana de ejemplo (Puga, 2012).

La estructura de la Red Bayesiana creada consta únicamente de 3 nodos, lo que significa que se están teniendo en cuenta 3 variables para dar solución al problema, que son justamente la cantidad de atributos que se tienen en la base de datos de prueba entre los cuáles se ha creado relaciones de causa-efecto. Ahora se procederá a definir los valores aceptados por cada una de las variables, para lo cual nuevamente se necesitará del apoyo de los expertos y de la base de datos (Ver Figura 3.4).

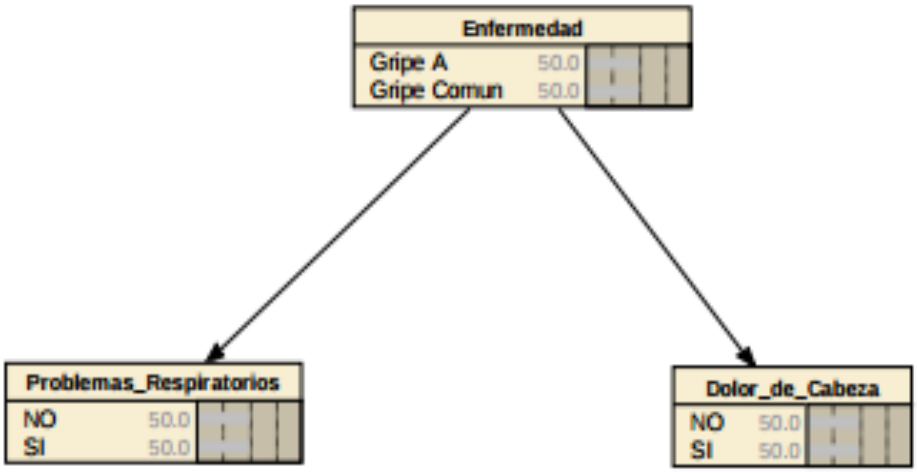


Figura 3.6. Estructura de las Red Bayesiana con los posibles valores de cada variable (Puga, 2012).

Una vez se tiene definida la estructura de la RB, se procederá con la Estimación Paramétrica, la cual será realizada de forma automática, es decir, se utilizará un algoritmo que procese el conjunto de datos mostrados en la Figura 3.6.

Como se había mencionado antes, el algoritmo pretende encontrar la estimación de máxima verosimilitud. Para fines prácticos del ejemplo, se aplicará un algoritmo basado en la

frecuencia relativa conjunta. La versión más sencilla del algoritmo, con una pequeña mejora a su versión original, introduciendo un factor de corrección quedaría de la siguiente manera:

$$p(x_i \mid x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)}) + 1}{n(x_{\pi(i)}) + \mid X_i \mid}$$

Figura 3.7. Fórmula utilizada para el cálculo de las probabilidades asociadas (Puga, 2012)

donde $n(x_{\pi(i)})$se refiere al número de casos que contiene la base de datos en los que la variable $X_{\pi(i)} = x_{\pi(i)}$, $n(x_i, x_{\pi(i)})$es el número de casos en que $X_i = x_i$y $X_{\pi(i)} = x_{\pi(i)}$y $|X_i|$es el número de estados que tiene la variable X_i.

Luego de aplicar la ecuación mostrada en la Figura 3.7 a cada una de las variables de la RB se da por terminado el proceso de Estimación Paramétrica dado que se tendrían calculadas las tablas de probabilidades asociadas a cada nodo como se muestra a continuación en la Figura 3.10.

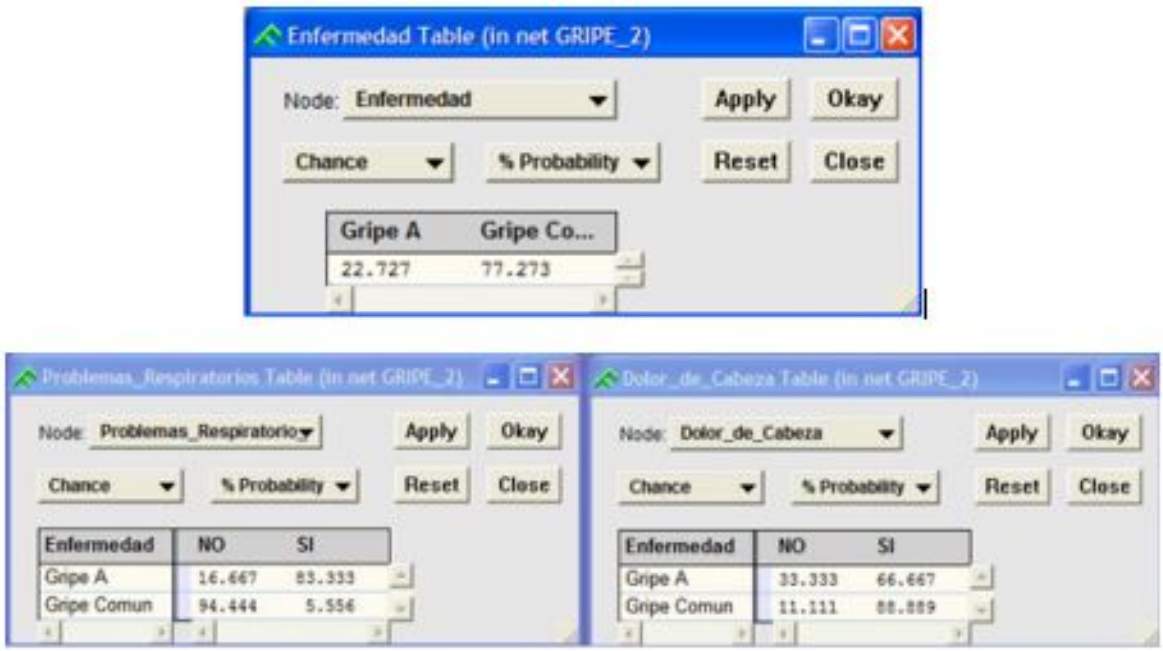


Figura 3.8. Tablas de probabilidad asociadas a la RB. Enfermedad (arriba), Problemas Respiratorios (izquierda), Dolor de cabeza (derecha).

4 CAPÍTULO IV: ANÁLISIS, DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

En el presente capítulo se describe paso a paso cómo se realizó la implementación del Sistema Experto Probabilístico basado en Redes Bayesianas utilizando la metodología descrita en el capítulo anterior.

4.1 SELECCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos utilizados proviene del repositorio público UCI. Los registros contienen información de 858 mujeres diferentes y fueron recolectados por el Hospital Universitario de Caracas, Venezuela. Si bien es cierto contiene un gran número de instancias, muchas pacientes decidieron no responder algunas de las preguntas por lo que el conjunto de datos tiene algunos registros con menor información que otros y viceversa.

4.2 TRATANMIENTO DE DATOS

Selección de Variables

Uno de los puntos más importantes para comenzar y llegar a implementar un modelo que requiere alimentarse de datos, es el de garantizar que el conjunto de datos es el adecuado para la técnica que se va a aplicar. Garantizar que el conjunto de datos está en buena forma significa que las probabilidades de que el experimento sea real y verosímil son bastante altas. Para este caso de estudio, se procedió analizando el conjunto completo de 858 instancias extraídas del repositorio UCI y eliminar aquellos registros que presentaban la mayor cantidad de atributos faltantes. Dado que uno de nuestros objetivos es hacer que nuestro modelo aprenda de un conjunto de datos, es importante que el conjunto del cual va a aprender sea lo más completo posible, un paciente del cuál no se conocen ni siquiera la mitad de las variables no añade mucho valor al proceso de aprendizaje del modelo probabilístico.

Como resultado nos quedamos con 322 registros en total para proseguir con la implementación del modelo basado en Redes Bayesianas.

La cantidad total de atributos en el conjunto de datos original es de 36, sin embargo, muchos de estos atributos también tienen valores faltantes, lo que significa que un gran número personas no recuerda o simplemente no desea revelar. Así como también hay atributos que

están tan fuertemente relacionados que pueden ser inclusive tratados como uno solo, o cuya relevancia está inmediatamente siendo considerada dentro de otra variable, lo cual nos hace prescindir del mismo.

Todo ello nos conlleva a limpiar el conjunto de datos de forma de quedarnos únicamente con aquello que sea necesario, relevante y se adapte a la metodología. Luego de haber realizado un análisis y prescindir de algunos atributos, el conjunto de datos final utilizado contiene un total de 322 registros y 15 variables cada uno, incluyendo la variable binaria target, la cual indica si el paciente fue diagnosticado con cáncer de cuello uterino o no. A continuación, la Tabla 4.1 muestra las variables 15 seleccionadas .

Variable	Descripción	Tipo
Edad	Edad actual de la paciente.	Continuo
Edad primera relación sexual	Edad de la primera relación sexual de la paciente.	Continuo
# Parejas sexuales	Número de parejas sexuales de la paciente.	Discreto
# Embarazos	Número de embarazos (Se toma en cuenta independiente si la paciente tuvo o no al bebé)	Discreto
ETS	Indica si la paciente alguna vez sufrió de alguna enfermad de transmisión sexual.	Discreto
# ETS	Número de enfermedades de transmisión sexual diagnosticadas previamente.	Discreto
DIU	Indica si la paciente alguna vez utilizó Dispositivo Intrauterino.	Discreto
Años con DIU	Número de años que la paciente lleva utilizando Dispositivo Intrauterino.	Continuo

Anticonceptivos Hormonales	Indica si la paciente alguna vez utilizó Anticonceptivos Hormonales.	Discreto
# Años con Anticonceptivos Hormonales	Número de años que la paciente lleva utilizando Anticonceptivos Hormonales.	Continuo
Fuma	Indica si la paciente alguna vez fumó	Discreto
# Años fumando	Número de años que la paciente lleva fumando.	Continuo
Test de Schiller	Resultado del Test Schiller, indica si el resultado fue positivo o negativo.	Discreto
Colposcopía	Resultado del examen de Colposcopía, indica si el resultado fue positivo o negativo.	Discreto
target	Diagnóstico de cáncer de cuello uterino.	Discreto

Tabla 4.1. Variables utilizadas en el trabajo de investigación (Elaboración Propia)

Discretización de Variables

Muchas de las variables son de tipo continuo, lo que significa que hace falta un proceso de Discretización de Variables. Por ejemplo, la variable Edad es una variable continua que necesita ser discretizada, para ello, los valores de la variable Edad se dividen en 4 estados, cada uno de ellos representando un intervalo, el primero de todas las personas cuya edad está entre 15 y 25 inclusive, el segundo de todas las personas cuya edad está entre 26 y 36 inclusive, el tercero entre 36 y 46 inclusive y finalmente el cuarto formado por todas las personas cuya edad es igual o mayor a 46. Igualmente, existen algunas variables que si bien es cierto son discretas, es posible dividir en intervalos debido a la gran cantidad de posibles valores que tienen a pesar de ser finitos, por ejemplo, la variable # Parejas Sexuales es de

tipo discreto, sin embargo podemos agrupar los valores de la siguiente forma: De 0 a 2 para personas que han tenido a lo mucho 1 pareja sexual en su vida, de 2 a 4 para personas que ha tenido 2 o 3 parejas sexuales, de 4 a 6 para personas que han tenido 4 o 5 parejas sexuales, y finalmente, de 6 a más para aquellas personas que han tenido un número alto de parejas sexuales a lo largo de su vida. La Tabla 4.2 muestra los posibles valores de todas las variables incluyendo los intervalos definidos para las variables de tipo continuo.

Variable	**Posibles Valores**
Edad	• De 15 a 26 • De 26 a 36 • De 36 a 46 • Mayores de 46
Edad primera relación sexual	• De 11 a 17 • De 17 a 25 • Mayores de 25
# Parejas sexuales	• De 0 a 2 • De 2 a 4 • De 4 a 6 • Más de 6
# Embarazos	• De 0 a 2 • De 2 a 4 • Más de 4
ETS	• 0 • 1
# ETS	• 0 • 1 • 2 • 3
DIU	• 0 • 1

Años con DIU	• De 0 a 3 • De 3 a 6 • De 6 a 9 • Más de 9
Anticonceptivos Hormonales	• 0 • 1
# Años con Anticonceptivos Hormonales	• De 0 a 3 • De 3 a 6 • De 6 a 9 • Más de 9
Fuma	• 0 • 1
# Años fumando	• De 0 a 3 • De 3 a 6 • De 6 a 9 • Más de 9
Test de Schiller	• 0 • 1
Colposcopía	• 0 • 1
target	• 0 • 1

Tabla 4.2. Variables utilizadas y sus posibles valores luego de la discretización. (Elaboración Propia)

4.3 CONSTRUCCIÓN DEL MODELO PROBABILÍSTICO

A continuación, procedemos con la construcción del modelo probabilístico haciendo uso de las variables discretas obtenidas en la fase anterior. Para poder construir la Red Bayesiana se empezará definiendo la topología, para luego proceder a hacer el cálculo de las probabilidades.

Estimación Estructural

La estimación estructural se hizo de forma manual, si bien es cierto existen mecanismos para construir la topología como el Aprendizaje de árboles, Aprendizaje de poliárboles y Aprendizaje de redes multiconectadas, estos son aproximaciones que pueden terminar dando resultados buenos para cierto conjunto de datos, pero la estructura no representa verdaderamente una relación real entre las variables. La idea principal del uso de las Redes Bayesianas es ofrecer un alto nivel de interpretabilidad, y es algo que solo se puede lograr con una estructura que verdaderamente represente el conocimiento sobre el área de dominio. Es por eso que, para el presente caso de estudio, la estructura topológica fue desarrollada con el apoyo del área de Ginecología y Obstetricia del Hospital Militar de la Fuerza Aérea del Perú y está basado en el conocimiento de expertos sobre el cáncer de cuello uterino.

Para proceder con la construcción del modelo, se utilizó la ayuda del software Netica 6.0.5. Netica es un software desarrollado por la empresa Norsys, y es una herramienta poderosa y fácil de utilizar al momento de trabajar con Redes Bayesianas. Netica ofrece una interfaz bastante intuitiva que permite al usuario interactuar con la Red Bayesiana gráficamente, así como también tiene muchos de los algoritmos de aprendizaje ya implementados, como el algoritmo EM y otros más como el Counting-Learning.

Con ayuda de Netica se pudo modelar el modelo de estimación estructural definido en conjunto de los expertos tal y como se puede observar en la Figura 4.1.

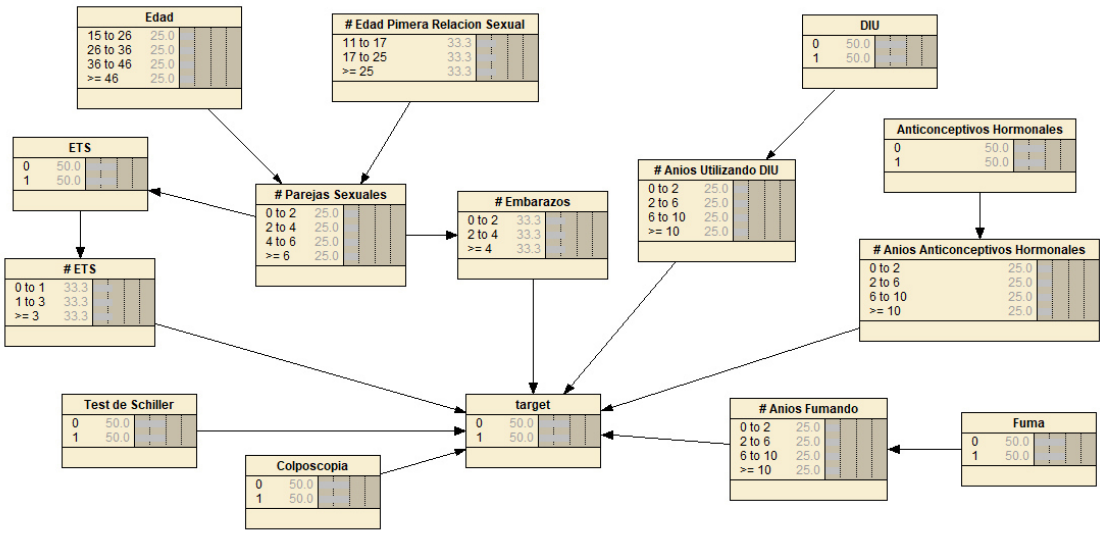


Figura 4.1.Topología de la Red Bayesiana en construcción (Elaboración Propia)

Estimación Paramétrica

Una se definió la estructura y topología de la Red Bayesiana, tuvo lugar el aprendizaje paramétrico. Nuevamente con la ayuda de Netica 6.0.5 se ejecutó el algoritmo EM para un conjunto total de 193 registros, los cuales representan el 60% del conjunto total de datos (322). La consola de Netica muestra el progreso y cada una de las iteraciones del algoritmo EM durante la fase de aprendizaje, para este caso de estudio fueron necesarias únicamente tres iteraciones para que el algoritmo EM detecte que ya no existen mayores cambios significativos respecto a las iteraciones previas. Se muestra en la Figura 4.2 que, durante la segunda iteración, las probabilidades variaron hasta en un 47.6519%, sin embargo, con los ajustes hechos durante las dos fases del algoritmo EM en la siguiente iteración, los cambios fueron reducidos a 0.0% por lo que el proceso de aprendizaje se considera finalizado.

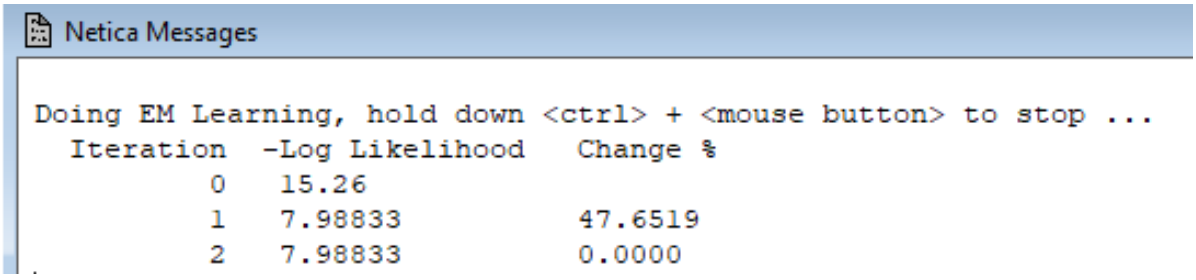


Figura 4.2. Consola de Netica con las iteraciones del algoritmo EM (Elaboración Propia)

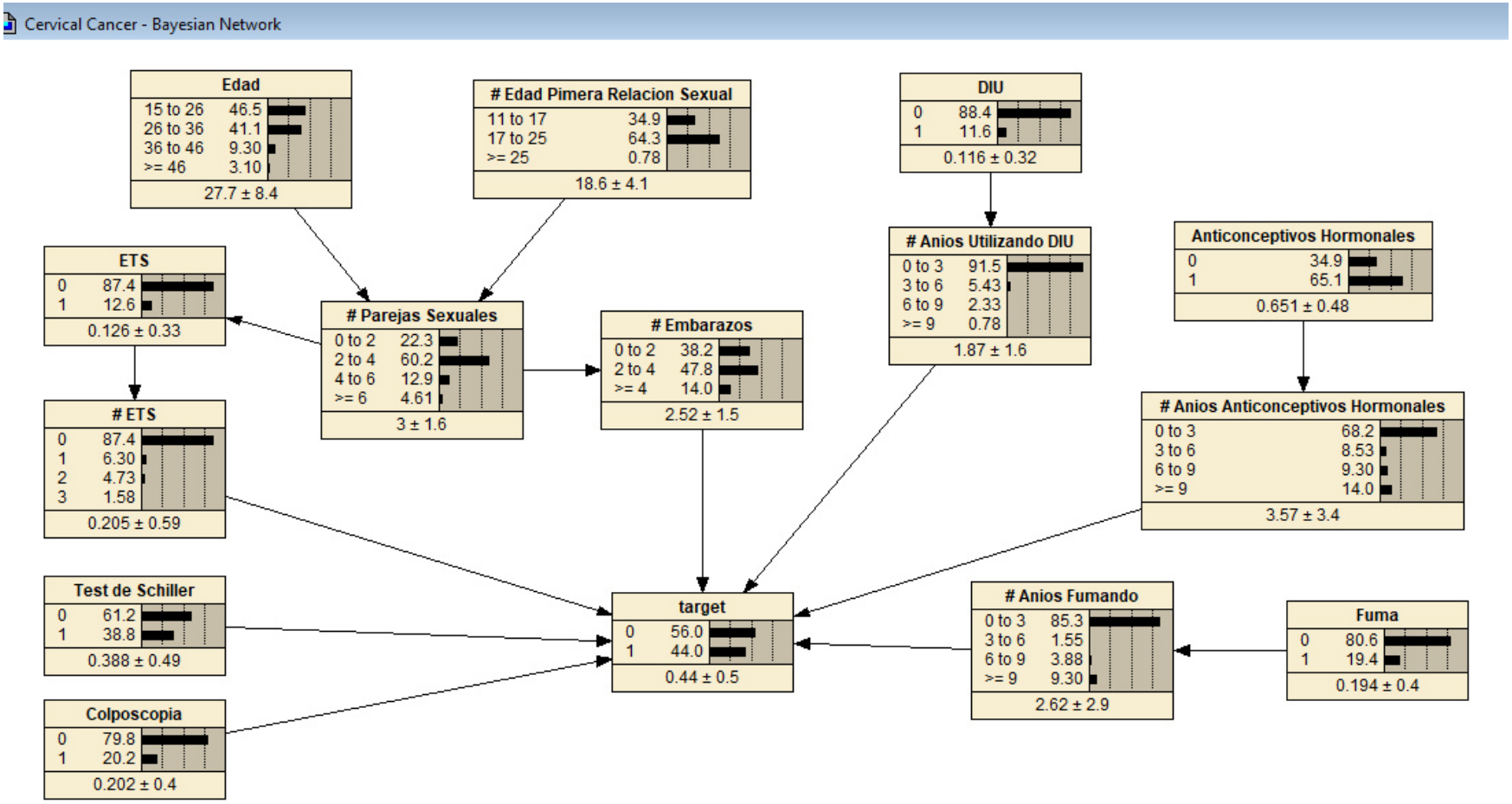


Figura 4.3. Estado de los nodos de la Red Bayesiana luego del Aprendizaje Paramétrico (Elaboración Propia).

Una vez este proceso ha terminado, se puede apreciar como la Red Bayesiana asigna una serie de valores a cada uno de los nodos (Ver la Figura 4.3). Estos valores representan las probabilidades marginales y condicionales utilizadas durante el proceso del algoritmo EM, dependiendo de si son nodos raíz o no respectivamente.

Ahora que se ejecutó la Estimación Paramétrica, cada nodo tiene una tabla de probabilidad asociada. Las tablas de probabilidad ayudan a inferir la probabilidad de una variable no observada basándose en la evidencia o valores previamente inferidos de los nodos padres. Netica, por supuesto, también ofrece una interfaz bastante elegante para trabajar con las tablas de probabilidades. La forma como las tablas están organizadas es la siguiente: la tabla se divide en dos secciones grandes, la sección izquierda, es sub dividida en columnas, una por cada nodo padre del nodo en análisis, mientras tanto, la sección derecha es dividida a su vez en columnas, una por cada valor que la variable del nodo en análisis puede tomar. Luego, en cada una de las celdas de la sección de la izquierda se colocan todas las combinaciones posibles para los diferentes valores que pueden tomar cada uno de las variables de los nodos padre, finalmente, cada una de estas combinaciones afectará a la probabilidad de cada uno de los diferentes estados o valores listados en la sección derecha (Ver desde la Figura 4.4 hasta la Figura 4.9).

_Parejas_Sexuales Table (in Bayes net Cervical_Cancer_Network)

Node: __Parejas_Sexuales | Apply | OK | Reset | Close | Chance | % Probability

Edad	# Edad Pimera Relacion Sexual	0 to 2	2 to 4	4 to 6	>= 6
15 to 26	11 to 17	14.286	67.857	17.857	3.57e-5
15 to 26	17 to 25	40.625	46.875	9.375	3.125
15 to 26	>= 25	25	25	25	25
26 to 36	11 to 17	23.077	38.461	23.077	15.385
26 to 36	17 to 25	10.256	76.923	7.692	5.128
26 to 36	>= 25	99.997	1.00e-3	1.00e-3	1.00e-3
36 to 46	11 to 17	3.33e-4	99.999	3.33e-4	3.33e-4
36 to 46	17 to 25	22.222	66.666	11.111	1.11e-4
36 to 46	>= 25	25	25	25	25
>= 46	11 to 17	1.00e-3	1.00e-3	99.997	1.00e-3
>= 46	17 to 25	3.33e-4	99.999	3.33e-4	3.33e-4
>= 46	>= 25	25	25	25	25

Figura 4.4. Tabla de probabilidad para la variable #Parejas Sexuales (Elaboración Propia)

_Embarazos Table (in Bayes net Cervical_Cancer_N...

Node: __Embarazos | Apply | OK | Reset | Close | Chance | % Probability

# Parejas Sexuales	0 to 2	2 to 4	>= 4
0 to 2	51.852	44.444	3.704
2 to 4	33.333	54.321	12.346
4 to 6	43.75	18.75	37.5
>= 6	20	60	20

Figura 4.5. Tabla de probabilidad para la variable #Embarazos (Elaboración Propia)

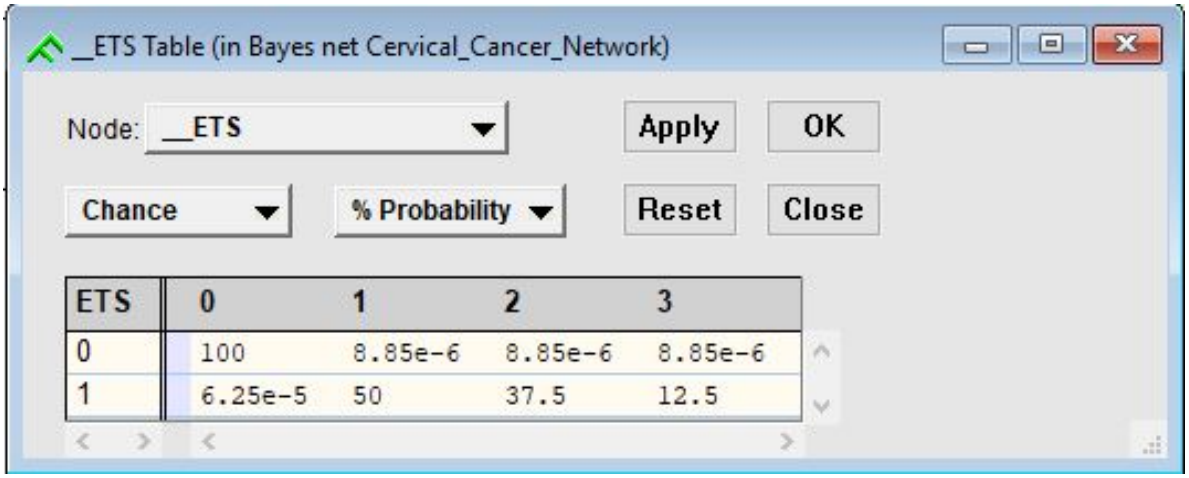


Figura 4.6. Tabla de probabilidad de la variable #ETS (Elaboración Propia).

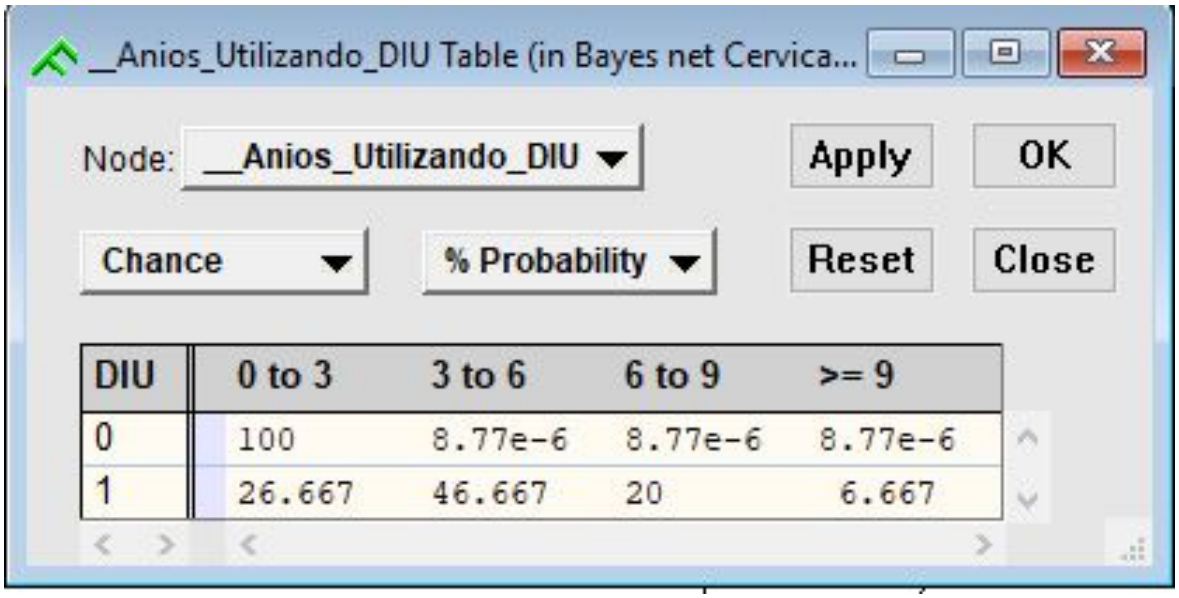
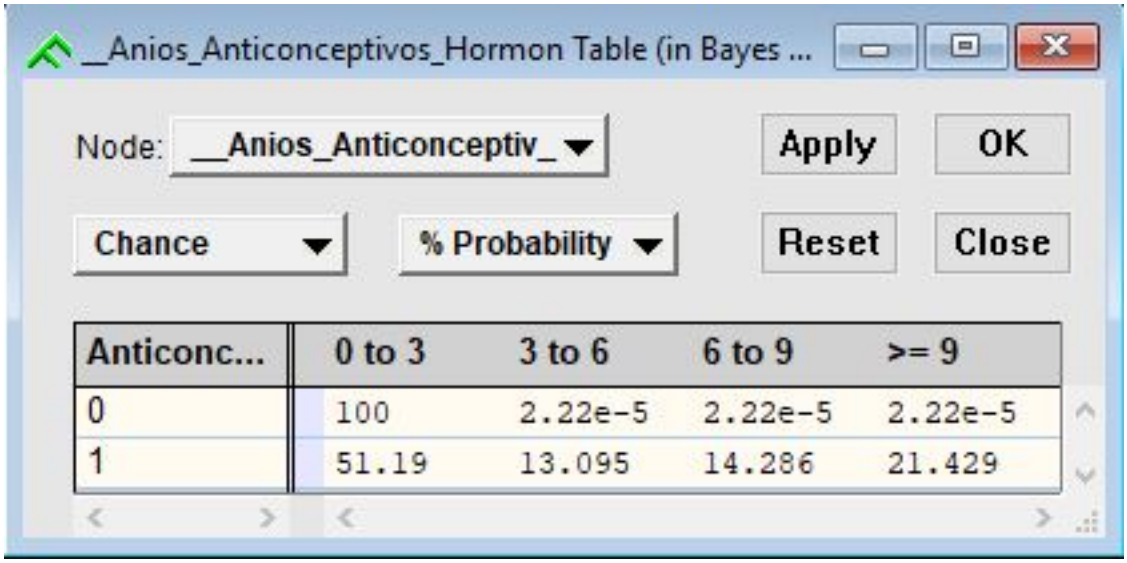
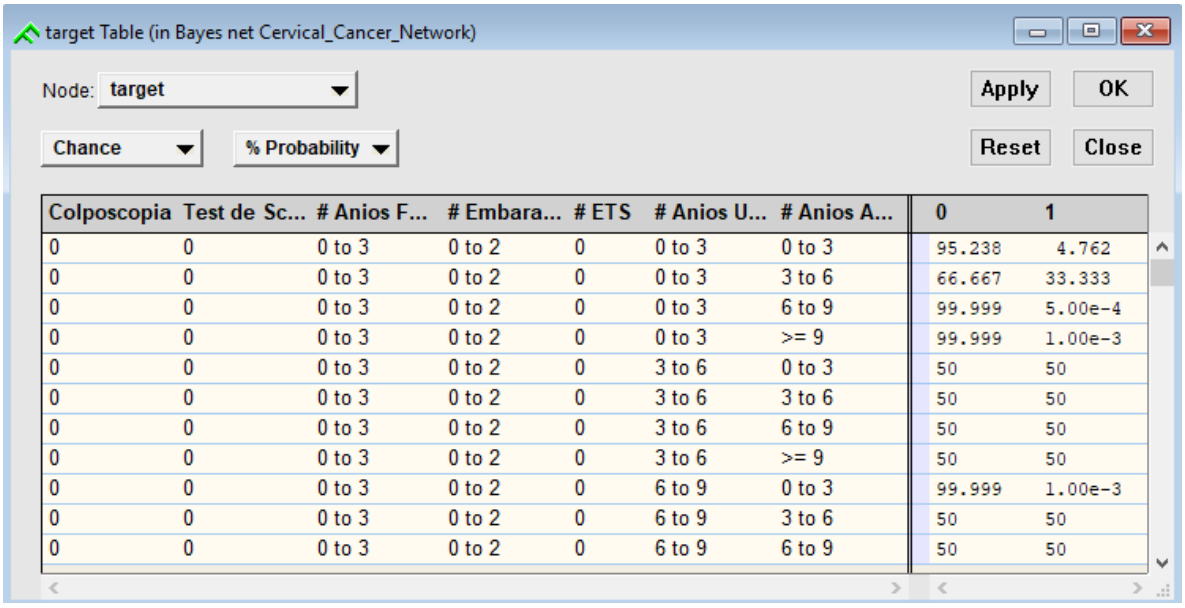


Figura 4.7. Tabla de probabilidad de la variable #Años utilizando DIU (Elaboración Propia).



Anticonc...	0 to 3	3 to 6	6 to 9	>= 9
0	100	2.22e-5	2.22e-5	2.22e-5
1	51.19	13.095	14.286	21.429

Figura 4.8. Tabla de probabilidad de la variable #Años utilizando anticonceptivos hormonales (Elaboración Propia).



Colposcopia	Test de Sc...	# Anios F...	# Embara...	# ETS	# Anios U...	# Anios A...	0	1
0	0	0 to 3	0 to 2	0	0 to 3	0 to 3	95.238	4.762
0	0	0 to 3	0 to 2	0	0 to 3	3 to 6	66.667	33.333
0	0	0 to 3	0 to 2	0	0 to 3	6 to 9	99.999	5.00e-4
0	0	0 to 3	0 to 2	0	0 to 3	>= 9	99.999	1.00e-3
0	0	0 to 3	0 to 2	0	3 to 6	0 to 3	50	50
0	0	0 to 3	0 to 2	0	3 to 6	3 to 6	50	50
0	0	0 to 3	0 to 2	0	3 to 6	6 to 9	50	50
0	0	0 to 3	0 to 2	0	3 to 6	>= 9	50	50
0	0	0 to 3	0 to 2	0	6 to 9	0 to 3	99.999	1.00e-3
0	0	0 to 3	0 to 2	0	6 to 9	3 to 6	50	50
0	0	0 to 3	0 to 2	0	6 to 9	6 to 9	50	50

Figura 4.9. Extracto de la tabla de probabilidad para la variable target (Elaboración Propia).

5 CAPÍTULO V: EXPERIMENTOS NUMÉRICOS

Se hizo una evaluación de 129 casos (40% del total de casos) para poder obtener métricas acerca del funcionamiento del modelo probabilístico propuesto. Para ello fue necesario 3 pasos importantes: primero, compilar la red; segundo, seleccionar una variable objetivo, la cual es justamente la variable que se tiene pensado evaluar; y finalmente proceder a ingresar los casos de prueba.

Tal y como se puede apreciar en la Figura 5.1, los casos de prueba se encuentran definidos en Excel utilizando un formato de CSV separado por comas. La primera fila contiene la descripción de los campos en cada una de las columnas, mientras que cada una de las filas siguientes representan una instancia nueva de caso de prueba. La Figura 5.1 muestra la definición de los primeros 30 casos, de un total de 129, utilizados para poner a prueba la eficiencia del modelo propuesto.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Edad	#Parejas Sexuales	#Edad Primera Relacion Sexual	#Embarazos	Fuma	#Anios Fumando	Anticonceptivos Hormonales	#Anios Anticonceptivos Hormonales	DIU	#Anios Utilizando DIU	ETS	#ETS	Colposcopia	Test de Schiller	target
2	51	3	17	6	1	34	0	0	1	7	0	0	1	1	1
3	40	1	18	1	0	0	1	0.25	0	0	1	2	0	1	1
4	40	1	20	2	0	0	1	15	0	0	0	0	1	1	1
5	37	2	18	2	0	0	0	0	1	5	1	1	0	1	1
6	37	3	19	3	1	12	1	13	0	0	0	0	0	1	1
7	38	2	15	4	0	0	1	16	0	0	0	0	0	1	1
8	33	1	29	2	0	0	1	0.5	0	0	0	0	0	1	1
9	35	5	11	2	1	15	1	14	0	0	0	0	1	1	1
10	38	3	18	4	0	0	1	10	1	2	0	0	0	1	1
11	30	1	13	3	1	22	0	0	0	0	1	1	1	1	1
12	28	5	14	4	0	0	1	3	0	0	0	0	0	0	1
13	28	2	17	1	0	0	1	9	0	0	0	0	1	1	1
14	28	2	19	2	0	0	1	0.42	1	3	1	2	1	1	1
15	25	3	17	2	0	0	1	0.08	0	0	0	0	1	1	1
16	30	1	17	3	0	0	1	4	0	0	0	0	1	1	1
17	29	3	17	3	1	10	1	6	0	0	0	0	1	1	1
18	24	2	18	4	0	0	0	0	0	0	1	1	1	1	1
19	22	3	17	1	0	0	0	0	0	0	1	2	1	1	1
20	22	2	15	2	1	5	1	6	0	0	0	0	0	1	1
21	21	1	17	2	0	0	1	3	0	0	0	0	1	1	1
22	22	3	15	1	0	0	1	1	0	0	0	0	1	1	1
23	20	1	19	1	0	0	1	0.25	0	0	1	1	0	0	1
24	29	2	18	4	0	0	0	0	0	0	0	0	0	1	1
25	22	4	16	1	0	0	1	0.5	0	0	1	1	0	1	1
26	21	4	15	1	0	0	0	0	0	0	0	0	0	1	1
27	20	2	14	4	1	3	1	0.5	0	0	0	0	0	0	1
28	19	2	15	3	0	0	1	0.58	0	0	0	0	1	1	1
29	26	3	15	1	0	0	1	0.33	1	3	0	0	1	1	1
30	20	3	17	2	0	0	1	0.25	0	0	0	0	0	0	1
31	35	2	17	3	0	0	1	1	0	0	0	0	1	1	1

Figura 5.1. 30 casos de prueba utilizados para la validación y resultados (Elaboración propia).

La forma en la que se realiza la evaluación es haciendo uso del algoritmo Belief Propagation. Este algoritmo define cómo el conocimiento generado a partir de la evidencia es propagado a través de toda la Red Bayesiana, de tal forma de alimentar cada uno de los nodos incluyendo al nodo que representa la variable objetivo.

El algoritmo utiliza lo que se conoce como 'intercambio de mensajes', mensajes de tipo λ y π, y sugiere que la probabilidad de un nodo dada cierta evidencia *E*, es $P(Bi \mid E) = \alpha\pi(Bi)\lambda(Bi)$. A lo largo de todo el proceso de inferencia, cada nodo debe guardar los valores de ambos vectores: λ y π además de las tablas de probabilidad.

El intercambio de mensajes consiste simplemente en que cada uno de los nodos de la Red Bayesiana deben enviar los mensajes correspondientes tanto a sus padres, como a sus hijos. Para ello se utilizan las ecuaciones mostradas a continuación en la Figura 5.2, la cual detalla la fórmula utilizada para enviar un mensaje de un nodo *B* a su nodo padre *A* y la Figura 5.3 que muestra cómo se procede para enviar un mensaje desde un nodo B hacia sus nodos hijos.

$$\lambda_B(Ai) = \sum_j P(B_j \mid A_i)\lambda(B_j)$$

Figura 5.2. Ecuación para envía mensajes ascendentes (Elaboración Propia).

$$\pi_k(Bi) = \alpha\pi(B_j)\prod_{l\neq k}\lambda_l(B_j)$$

Figura 5.3. Ecuación para enviar mensajes descendentes (Elaboración Propia).

Con el apoyo de Netica, se procede primero a compilar la red, luego se elige la variable cuyo nombre es 'target' para ser usada como variable objetivo en la prueba, y finalmente se procede a ingresar el archivo CSV con los 129 casos de prueba.

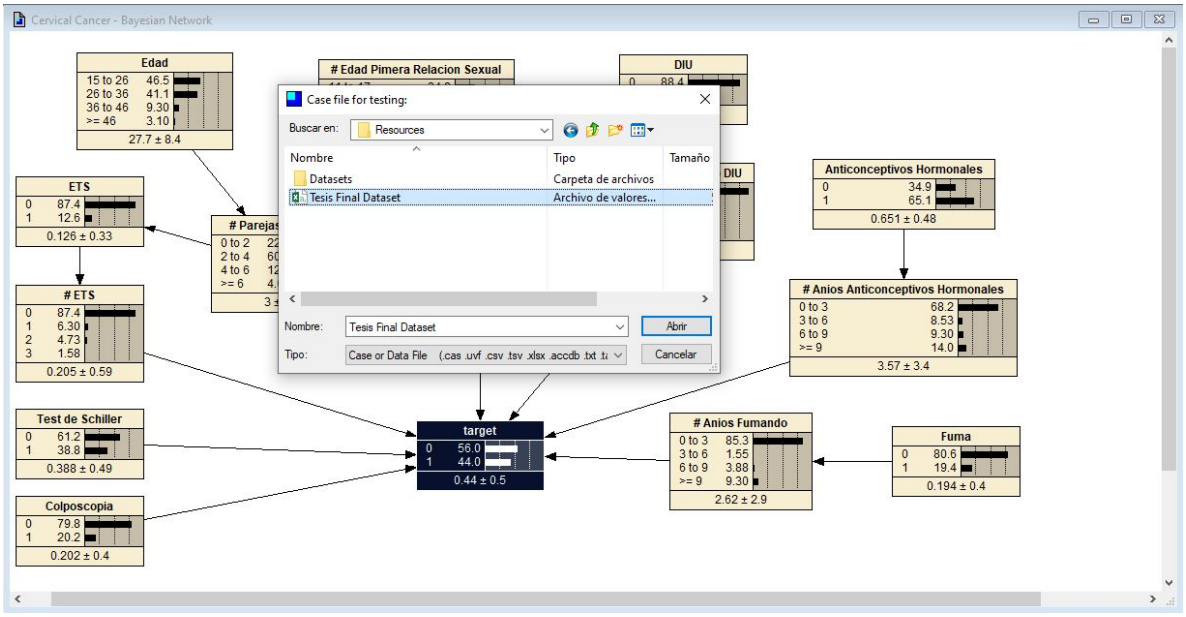


Figura 5.4. Ingresando los registros a ser utilizados en la validación (Elaboración Propia).

A continuación en la Tabla 5.1, se muestra la matriz de confusión resultante del proceso de validación.

Resultado Obtenido		
0	**1**	**Resultado Esperado**
73	1	**0**
4	51	**1**

Tabla 5.1. Matriz de confusión de la predicción (Elaboración Propia)

De la matriz de confusión mostrada, podemos deducir lo siguiente:

- De un total de 74 casos con resultado esperado negativo, 73 fueron evaluados correctamente como negativos, lo cual hace un total de 73 verdaderos negativos.
- De un total de 55 casos con resultado esperado positivo, 51 fueron evaluados correctamente como positivos, lo cual hace un total de 51 verdaderos positivos.
- De los 74 casos con resultado esperado negativo, 1 fue evaluado como positivo, es decir, existe 1 falso negativo.
- De los 55 casos con resultado positivo, 4 fueron evaluados como negativo, es decir, existe 4 falsos positivos.
- De un total de 129 predicciones, 124 (verdaderos negativos más verdaderos positivos) fueron hechas correctamente.

Para calcular el margen de error del modelo predictivo procedemos a dividir la cantidad de predicciones correctas entre el total de predicciones y restarle este valor a la unidad.

$$124 / 129 = 0.96124$$

$$Error = 1 - 0.96124 = 0.03876 = 3.3876 \times 1 / 100$$

$$\boldsymbol{Error = 3.876\ \%}$$

Haciendo uso de la Matriz de confusión creada, se procede a evaluar el método propuesto en términos de precisión, exactitud, sensibilidad y especificidad.

El primer atributo a calcular es la precisión del modelo probabilístico. La precisión representa la dispersión (diferencia de varias medidas realizadas bajo las mismas condicionas contra un

valor esperado) del conjunto de valores obtenidos en la predicción, mientras menos dispersión, mayor es la precisión, y para calcularla, es necesario dividir la cantidad de predicciones cuyo resultado fue igual al esperado entre el total de predicciones, algo que se hizo previamente para poder calcular el margen de error, motivo por el que se concluye fácilmente que la precisión del modelo es igual a 0.96124.

$$Precisión = 0.96124 = 96.124 \times 1 / 100$$

$$\boldsymbol{Precisión = 96.124\ \%}$$

El siguiente atributo a evaluar es la exactitud. La exactitud indica lo cerca que está el resultado de una medición del valor esperado y está representada por la división de los verdaderos positivos entre todos los resultados positivos (verdaderos y falsos positivos).

$$Exactitud = 51 / (51 + 4)$$

$$Exactitud = 0.92727 = 92.727 \times 1 / 100$$

$$\boldsymbol{Exactitud = 92.727\ \%}$$

Por otro lado, la sensibilidad representa la capacidad del modelo predictivo para predecir un valor positivo en caso realmente lo sea, en otras palabras, la capacidad de predecir resultados positivos correctamente. El cálculo de la sensibilidad viene dado por la división de la cantidad de verdaderos positivos con la cantidad total de casos en los que el resultado esperado era positivo (verdaderos positivos y falsos negativos).

$$Sensibilidad = 51 / (51 + 1)$$

$$Sensibilidad = 0.98076 = 98\text{-}076 \times 1 / 100$$

$$\boldsymbol{Sensibilidad = 98.076\ \%}$$

La especificidad funciona de manera análoga a la sensibilidad, pero para los casos negativos, es decir, indica la capacidad de predecir resultados negativos correctamente. El cálculo se

realiza al dividir la cantidad de verdaderos negativos con el total de casos cuyo resultado esperado era negativo (verdaderos negativos y falsos positivos).

$$Especificidad = 73 / 73 + 4$$

$$Especificidad = 0.94805 = 0.94805 \times 1 / 100$$

$$\boldsymbol{Especificidad = 94.805\ \%}$$

Finalmente, la Tabla 5.2 resume los resultados obtenidos por el modelo probabilístico:

Precisión	Exactitud	Sensibilidad	Especificidad
96.124%	92.727%	98.076%	94.805%

Tabla 5.2. Características obtenidas a partir de la Matriz de Confusión (Elaboración propia).

En base a los resultados obtenidos, se puede afirmar que el modelo predictivo propuesto es confiable y tiene un alto rendimiento, especialmente en la predicción de casos positivos qué es donde el modelo predictivo alcanzó mejores resultados.

6 CAPÍTULO VI: CONCLUSIONES Y TRABAJOS FUTUROS

- Se implementó un sistema probabilístico basado en Redes Bayesianas capaz de clasificar con una tasa de éxito de 96% a personas diagnosticadas con cáncer de cuello uterino. Además de ello, gracias a la versatilidad y transparencia que ofrecen las Redes Bayesianas, se puede analizar las probabilidades de cada una de las variables. Las Redes Bayesianas has demostrado poder alcanzar un rendimiento bastante alto ofreciendo total transparencia sobre el proceso de inferencia, algo que no pasa con muchas otras técnicas que por lo general ofrecen un resultado, pero sin tener en cuenta la incertidumbre existente y más aún, sin poder explicar el cómo se llega a esa conclusión.
- Los resultados obtenidos por el modelo probabilístico basado en Redes Bayesianas del presente trabajo de investigación lograron superar largamente, en términos de precisión, a todos los experimentos realizados entre el año 2005 y 2015 con una cantidad de atributos y número de registros similares que fueron mencionados por la Universidad de Ciencias Médicas, Irán (Ver Figura 3.2).
- Uno de los trabajos más recientes, publicado en la Revista Internacional de Ciencias de la Computación Avanzada mencionaba una lista de trabajos previamente implementados en donde no fueron utilizadas las Redes Bayesianas. Haciendo una pequeña comparación entre los resultados obtenidos en esta investigación con los resultados de la investigación desarrollada por dos universidades de Pakistán (Mahboob, Milhan, Iqbal, Wahab, & Mushtaq, 2019), se puede ver que las Redes Bayesianas superan a todos los métodos mencionados en términos de precisión, lo cual es un muy buen indicador e incentiva a seguir aplicando esta técnica a diferentes casos de estudios relacionados con predicción.
- A pesar de los muy buenos resultados, no hay que dejar de recordar que el conjunto de datos utilizado fue de 322 instancias y 15 atributos. Sería ideal poder realizar más experimentos con esta técnica sobre conjuntos de datos muchísimo más grandes, que incluyan miles de instancias y muchos más atributos. Hoy en día el principal obstáculo es la recolección de datos, ya que se requiere de bastante tiempo, paciencia, perseverancia y, sobre todo, colaboración por parte de las instituciones de salud para proveer la mayor cantidad de datos posibles de las pacientes atendidas para fines académicos.
- El modelo probabilístico propuesto maximiza la cantidad de verdaderos positivos alcanzando una taza del 98% al momento de identificar pacientes con riesgo de cáncer,

esto es muy importante sobre todo dentro del área de dominio de la Medicina, ya que cometer un error al predecir un caso verdadero negativo no tiene el mismo impacto que comeré un error al predecir un caso verdadero positivo, es por ello que siempre se busca minimizar esto último, algo que el modelo propuesto basado en Redes Bayesianas logra al maximizar la eficiencia de identificar verdaderos positivos.

7 REFERENCIAS BIBLIOGRÁFICAS

Bountris, P., Tsirmpas, C., Koutsouris, D., & Haritou, M. (2014). Bayesian Networks to Support the Management of Patients with ASCUS/LSIL Pap Tests. *Conference: 4th International Conference on Wireless Mobile Communication and Healthcare (MOBIHEALTH)*.

Carmona Sánchez, E. (2014, Julio 11). Tutorial sobre Máquinas de Vectores de Soporte (SVM). Madrid, España: Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED).

Carvalho, P. (2015, Diciembre 9). *An Aspiring Rationalist's Ramble*. Retrieved from https://rationalistramble.wordpress.com/2015/12/09/bayesian-networks/

Castillo, E., Gutiérrez, J., & Hadi, A. (1997). *Expert systems and probabilistic network models.* Springer-Verlag New York. doi:10.1007/978-1-4612-2270-5

Centros para el Control y la Prevención de Enfermedades. (2014, Noviembre 5). *Millones de mujeres en los EE. UU. no se están haciendo las pruebas de detección de cáncer de cuello uterino*. Retrieved from https://www.cdc.gov/spanish/mediosdecomunicacion/comunicados/p_vs_cancer_cuello_uterino_110514.html

Chang, C.-C., Cheng, S.-L., Lu, C.-J., & Liao, K.-H. (2013). Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification. *International Journal of Machine Learning and Computing vol. 3, no. 1*, 75-78.

Correa Valencia, M., Bielza, C., Pamies Teixeira, J., & R. Alique, J. (2008). Redes Bayesianas vs Redes Neuronales en modelos para la predicción del acabado superficial. *XVII Congreso de Máquinas-Herramienta y Tecnologías de Fabricación*.

DeWilde, B. (2012, 10 26). *Burton DeWilde*. Retrieved from Classification of Hand-written Digits (3): http://bdewilde.github.io/blog/blogger/2012/10/26/classification-of-hand-written-digits-3/

Díez, F. J. (1998). *Aplicaciones de los modelos gráficos probabilistas en medicina.* Retrieved from RUIdeRA: http://hdl.handle.net/10578/6174

Gómez Fernández, J. M. (2013, Abril 19). *Pseudocódigo y algunas herramientas de Minería de Datos*. Retrieved from http://id3gocuteam.blogspot.com/2013/04/pseudocodigo-y-algunas-herramientas-de.html

Hatcher, M. (2014, Julio 22). *The Fact Machine*. Retrieved from http://www.thefactmachine.com/random-forests/

Hu, L., Bell, D., Antani, S., Xue, Z., Yu, K., Horning, M., . . . Schiffman, M. (2019). An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. Retrieved from https://doi.org/10.1093/jnci/djy225

INEM. (2013). *Guía de práctica clínica de cáncer de cuello uterino.* Instituto Nacional de Enfermedades Neoplásicas, Dpto de Oncología Médica, Perú.

Introducción a las redes neuronales artificiales. (n.d.). Retrieved from Introducción a las redes neuronales artificiales: http://www.rna.50webs.com/tutorial/RNA_intro.html

Juárez Ibujes, M. O. (2016, Junio 9). *Probabilidad Total y Teorema de Bayes*. Retrieved from https://www.monografias.com/trabajos89/probabilidad-total-y-teorema-bayes/probabilidad-total-y-teorema-bayes.shtml

Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. *Acta Informatica Medica*. doi:10.5455/aim.2016.24.364-369

Liang, M., Zheng, G., Huang, X., Milledge, G., & Tokuta, A. (2013). Identification of Abnormal Cervical Regions from Colposcopy Image Sequences. *21st International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 130-136.

Mahboob, T., Milhan, M., Iqbal, M., Wahab, A., & Mushtaq, M. (2019). Cervical Cancer Prediction through Different Screening Methods using Data Mining. *International Journal of Advanced Computer Science and Applications, 10*, 388-396.

Matsuo, K., Purushotham, S., Jiang, B., S. Mandelbaum, R., Takiuchi, T., Liu, Y., & Roman, L. (2019). Survival outcome prediction in cervical cancer: Cox models versus deep-

learning model. *American Journal of Obstetrics and Gynecology, 220*, 381.e1-381.e14.

McCullock, J. (2013). *Mnemosyne Studio*. Retrieved from k-means Introduction: http://www.mnemstudio.org/clustering-k-means-introduction.htm

Onísko, A., Druzdzel, M. J., & Austin, R. M. (2009). Application of Dynamic Bayesian Networks to. *Intelligent Information Systems 9999*, 1-10.

Organización Mundial de la Salud. (2019). *Papilomavirus humanos (PVH) y cáncer cervicouterino*. Retrieved from https://www.who.int/es/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

Pereira, R. T., & Chamorro, M. C. (2012). La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. *Universidad y Salud vol.14 no. 2*.

Puga, J. L. (2012). Cómo construir y validar Redes Bayesianas con Netica. *Revista electrónica de Metodología Aplicada, 17*(1), 1-17.

QuarkGluon Ltd. (2018, Junio 3). *Beginner's Guide to Decision Trees for Supervised Machine Learning*. Retrieved from https://www.quantstart.com/articles/Beginners-Guide-to-Decision-Trees-for-Supervised-Machine-Learning

Ramachandran, P., Girija, N., & Bhuvaneswari, T. (2014). Early Detection and Prevention of Cancer using Data Mining Techniques. *International Journal of Computer Applications, 97*(13), 48-53.

Rubio, F., Martínez-Gómez, J., Flores, M. J., & Puerta, J. M. (2016). Comparison between Bayesian network classifiers and SVMs for. *Expert Systems With Applications*.

Rulequest. (2017, Febrero). *Is See5/C5.0 Better Than C4.5?* Retrieved from https://rulequest.com/see5-comparison.html

Sanchez, L. Y. (2012). Desarrollo de un Sistema Experto sobre web para un diagnóstico temprano de Cáncer de Cuello Uterino en la Clínica Maternidad "Belén" - Chiclayo.

Tesis presentada en la Facultad de Ingeniería de la Universidad Católica Santo Toribio de Mogrovejo.

Sancho Caparrini, F. (2017, Octubre 15). *Sistemas Basados en Reglas*. Retrieved from http://www.cs.us.es/~fsancho/?e=103

Sancho Caparrini, F. (2018, Enero 5). *Aprendizaje Inductivo: Árboles de Decisión*. Retrieved from http://www.cs.us.es/~fsancho/?e=104

Sato, M., Horie, K., Hara, A., Miyamoto, Y., Kurihara, K., Tomio, K., & Yokota, H. (2018). Application of deep learning to the classification of images from colposcopy. *Oncol Lett, 15*, 3518-3523.

Sayad, S. (2012, Noviembre 16). *Decision Tree - Classification*. Retrieved from http://www.saedsayad.com/decision_tree.htm

Soumya, M., Sneha, K., & Arunvinodh, C. (2016). Cervical Cancer Detection and Classification using Texture Analysis. *Biomedical and Pharmacology Journal, 9*(2).

Suryatenggara, J., Ane, B. K., Pandjaitan, M., & Steinberg, W. (2009). Pattern recognition on 2D cervical cytological digital images for early detection of cervix cancer. *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*.

Trevino, A. (2016, December 06). *Learn Data Science, Machine Learning*. Retrieved from Introduction to K-means Clustering: https://www.datascience.com/blog/k-means-clustering

Vidya, R., & Nasira, G. M. (2016). Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns. *Indian Journal of Science & Technology, 9*(30).

Wei, L., Gan, Q., & Ji, T. (2017). Cervical cancer histology image identification method based on texture and lesion area features. *Computer Assisted Surgery, 22:sup1*, 186-199. doi:10.1080/24699322.2017.1389397

Will. (2016, Febrero 27). *Decision Tree Flavors: Gini Index and Information Gain*. Retrieved from http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/

World Health Organization. (2013). *WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention.*

Xu, T., Zhang, H., Xin, C., Kim, E., Long, L., Xue, Z., . . . Huang, X. (2017). Multi-feature based benchmark for cervical dysplasia classification evaluation. *Pattern Recognition, 63*, 468-475.